A Study on Provenance Description of Metadata Schemas for Longevity of Metadata in Networked Information Environment

A Study on Provenance Description of Metadata Schemas for Longevity of Metadata in Networked Information Environment

March 2018

LI CHUNQIU

A Study on Provenance Description of Metadata Schemas for Longevity of Metadata in Networked Information Environment

LI CHUNQIU

Graduate School of Library, Information and Media Studies

University of Tsukuba

March 2018

A Study on Provenance Description of Metadata Schemas for Longevity of Metadata in Networked Information Environment

Abstract

It is widely recognized that longevity of digital resources is crucial in our networked information society and that metadata plays key roles in keeping digital resources usable over time. Metadata longevity must be ensured for longevity of the preserved resources. There are well-known standards for digital preservation, such as Open Archival Information System (OAIS) reference model and Preservation Metadata: Implementation Strategies (PREMIS). These standards define metadata models for digital preservation. However, they do not provide any models or guidelines to keep metadata interpretable over time. In other words, they do not include maintenance issues of metadata schemas that define representation schemes, structural features and semantics of metadata. More importantly, the longevity issues of metadata schemas are still largely unexplored. The author initiated the study presented in this dissertation from this basic standing point.

This study has three fundamental concepts as its basis: the formal description of metadata and their schemas suitable to the Semantic Web, Dublin Core Application Profiles (DCAP) as the basic framework of metadata schemas, and provenance description of metadata schemas.

The information environment of metadata has changed along with the progress of the Web. In the conventional information environment, metadata is stored in a database and accessed via an interface to the database. In the up-to-date Semantic Web environment, metadata and their schemas are defined in formal description schemes such as Resource Description Framework (RDF) and Web Ontology Language (OWL), and they can be transferred and shared as a digital object. Therefore, we need to develop technologies suitable to the Semantic Web environment for the longevity of metadata schemas.

Singapore Framework for DCAP defined by the Dublin Core Metadata Initiative is a well-known framework of metadata schemas. DCAP defines the components of a metadata schema for an application and related components such as metadata vocabularies for metadata interoperability. The DCAP explicitly separates semantic definitions of metadata terms and structural definitions of metadata constraints. The Singapore Framework is a layered model in which application specific features such as structural constraints and implementation syntax are defined in a layer above application neutral features which include definitions of metadata terms. The structural definitions of metadata constraints are formally described as Description Set Profiles (DSP) of metadata application profiles, and the semantic definitions of metadata terms are provided in metadata

vocabularies. This clear separation suggests that long-term maintenance of metadata application profiles and metadata vocabularies are the key issues for metadata longevity.

In the long-run, requirements and technologies for metadata may change, which may cause either or both structural and semantic changes in metadata schemas. Those changes may cause inconsistency in the use of metadata, which is a significant risk for the long-term use of digital resources. Therefore, both structural and semantic changes in metadata schemas should be consistently recorded and maintained over time. This study focuses on provenance description of metadata schemas that tracks structural and semantic changes in metadata schemas for long-term maintenance of metadata schemas.

The author has learned from OAIS and PREMIS that provenance information is important for longevity of digital resources and that provenance of metadata schema is required for long-term use of metadata. In general, provenance of a metadata schema includes descriptions about the change history of the metadata schema, agents responsible for its custody, key events that occurred over its lifecycle, and other information related to the creation, management, and preservation of the metadata schema. However, through literature review the author has learned that the existing provenance models are not designed for describing provenance of metadata schemas, in particular for tracking their change history. Therefore, in this study, the author aims to define provenance description models for tracking both structural and semantic changes in metadata schemas.

Based on the analysis of demands for long-term maintenance of metadata schemas given above, the author has developed two basic models to describe provenance of metadata schemas – one for DSP and the other for metadata vocabularies. The proposed models have their bases on the provenance description standard PROV defined by the Provenance Working Group at the World Wide Web Consortium (W3C) for description of provenance in the Web environment. W3C PROV standard is selected as a base to formally describe provenance of metadata schemas due to its strong extendibility and interchangeability of provenance description following PROV in heterogeneous environments. The machine-processable provenance description can be provided using W3C PROV Data Model (PROV-DM) and PROV Ontology (PROV-O). Thus, the models proposed in this study are aimed for the formal provenance description of metadata schemas that conforms to the requirements of Semantic Web environment.

In the early stage of this study, the author experimentally developed provenance descriptions of metadata schemas through a combination of PREMIS and PROV. The author provided provenance description examples using PROV-O and PREMIS OWL Ontology. Then the author applied W3C PROV to describe provenance of metadata application profiles and metadata vocabularies, respectively. The author tried to properly record the revision history of structural

constraints defined in metadata application profiles and definitions of metadata terms as formal provenance descriptions for the consistent maintenance of metadata. As Entity and Activity defined in PROV-DM are the key classes to describe provenance, the author defined a set of Entities and Activities as their sub-classes to track changes in metadata application profiles and metadata vocabularies, respectively. The author finally proposed two provenance models, i.e., DSP-PROV model for tracking the structural changes of metadata constraints in metadata application profiles and Vocab-PROV model for tracking semantic changes of metadata terms in metadata vocabularies.

DSP-PROV enables tracking revision, deletion and addition of description templates, statement templates and structural constraints defined in DSP. The author applied DSP-PROV to Digital Public Library of America Metadata Application Profile (DPLA MAP) as a case study to show the advantage of the model against semi-formal provenance description in change logs of DPLA MAP.

Vocab-PROV enables effective and automated tracking of change history of metadata vocabularies. The author defined a few primitive change types of metadata terms with functions to track the revision, deletion, addition, replacement of a metadata term and its definitions. The author also provided examples of provenance description in RDF graphs to show Vocab-PROV.

In this study, the author examined limitations and implications of DSP-PROV and Vocab-PROV. In practice, structural and semantic changes in metadata schemas may be more complicated than the experiments conducted in this study because of complexity of metadata schemas. However, the author considers that the proposed models in this study serve to track provenance of metadata schemas, help long-term maintenance of metadata schemas, extend functions of metadata registries, and audit errors in metadata mapping.

From this study, the author has learned that: (1) Keeping metadata consistently interpretable, not only by humans but also by machines, is a fundamental requirement of metadata longevity on the Web, and metadata longevity requires long-term maintenance of metadata schemas; (2) Long-term maintenance of metadata application profiles and metadata vocabularies are important issues for long-term maintenance of metadata schemas; (3) Provenance description should be machine-readable, interoperable and traceable for provenance interchange in the Web environment; (4) The structural and semantic changes in metadata schemas can be examined separately and these changes may synchronously happen; (5) Formal provenance descriptions following Web standards hold advantages over semi-formal provenance description written in a natural language.

ネットワーク情報環境におけるメタデータの長期利用性向上のためのメタデータスキーマの来歴記述に関する研究

概要

様々な情報資源がデジタル形式で作られるネットワーク情報化社会において、デジタルリソースの長期利用性が重要な課題であること、そしてデジタルリソースの長期利用を可能にするには、メタデータの長期利用性を保証することが求められることが広く知られている。例えば、デジタル保存のための国際標準としてよく知られる Open Archival Information System (OAIS)参照モデルや Preservation Metadata: Implementation Strategies (PREMIS)は、いずれもデジタル保存のためのメタデータの標準モデルを定めている。こうしたことからも、デジタル保存にはメタデータの長期維持管理の必要性が理解できる。メタデータの長期維持管理には、メタデータの実体（メタデータインスタンス）の維持管理のみならず、そのスキーマや関連規則などの長期維持管理が必要である。しかしながら、従来のデジタル保存の研究ではメタデータの長期利用性の維持、特にメタデータスキーマの長期に渡る一貫した維持管理の問題はあまり扱われてこなかった。

こうした背景の下で、筆者はメタデータスキーマの長期維持管理に焦点を当てた研究を進めた。本研究を進めるにあたり、筆者は、Semantic Web 環境に適したメタデータ及びスキーマの形式的記述、メタデータスキーマの基本フレームワークとなるダブリンコア・アプリケーションプロファイル(Dublin Core Application Profiles (DCAP)) とメタデータスキーマの来歴記述の 3 つの観点を基盤として本研究を進めることとした。以下の段落にこれらの観点について示す。

Web の発展によりメタデータの作成・利用の環境は大きく変化した。このことはメタデータの維持管理にも影響を与えている。例えば、従来の情報環境におけるメタデータはデータベースに蓄積、利用される。これに対し、現在の Web 環境におけるメタデータはデータベースに蓄積されるのみならず、デジタルオブジェクトとして Web 上で送受・共有されるようになった。特に、Semantic Web 環境では、メタデータの相互運用性に加えてコンピュータによる解釈可能性を高めるために、RDF や OWL といった形式的記述を用いてメタデータスキーマを定義することが行われる。したがって、メタデータの長期維持

管理においても、こうした Semantic Web 環境指向の形式的表現を基盤とすることが求められると考えられる。

DCAP は、メタデータ記述に用いられる語彙の意味的な定義とメタデータの構造的な制約を階層的に分けることでメタデータの相互運用性に関する要件を明示的に示している。DCAP の Web 環境における構造を明確に定義したシンガポールフレームワークは、応用向けのメタデータスキーマを構成する記述セットプロファイル(Description Set Profile (DSP))やドメインモデル(Domain Model)他全部で 5 つの構成要素を持つ。シンガポールフレームワークはメタデータ語彙を定義する階層の上に記述セットプロファイル等のメタデータスキーマ構成要素を定義する階層を定義している。すなわち、応用ごとに決まる構造的性質に関わる定義と、応用間で共通に利用可能なメタデータの記述のための語句（メタデータターム）とその集まりであるメタデータ語彙の定義を明確に分離することで、応用間でのメタデータ語彙の共有を可能にする一方、応用ごとにメタデータの構造的制約を決めることができるようにしている。

Semantic Web 環境では、応用向けに定義されるメタデータアプリケーションプロファイルや応用間に共通に定義されるメタデータ語彙も、デジタルオブジェクトとして Web 上で流通する。メタデータインスタンスと同様にメタデータスキーマも、相互運用性、機械解釈可能性を保持しながら長期に渡って維持する必要がある。そこで、筆者は、構造定義である記述セットプロファイルと語句の意味を決めるメタデータ語彙定義に焦点を当ててメタデータスキーマの長期維持管理について研究することが重要であると考えた。

長期に渡るメタデータの維持管理においては、長い時間の経過によって、メタデータの構造定義やメタデータ語彙の意味定義が変化し、その変化によってメタデータの解釈に矛盾が起きる可能性がある。これは、デジタルリソースの長期利用の潜在的なリスクになる。そのため、メタデータの構造的性質とメタデータ語彙の定義を適切に記録し、維持管理する必要がある。そのため、本研究では、メタデータスキーマの構造的性質とメタデータ語彙の変化を追跡するために利用することのできる情報、すなわち来歴情報(Provenance)に注目することとした。

筆者は OAIS と PREMIS から、デジタルリソースの長期利用のための来歴情報の重要性を学び、そこからデジタルリソースの長期利用を支えるメタデータの長期の維持管

理のためのメタデータスキーマの来歴記述の必要性に気づいた。一般的に、メタデータスキーマの来歴には、メタデータスキーマの変更履歴、保管等の責任を負うエージェント、変化のキーとなるイベント、メタデータスキーマの作成・管理・保存等に関する情報を記述する。筆者は、文献調査から、既存の来歴記述モデルをそのままメタデータスキーマ来歴記述には適用できないという知見を得た。そこで、本研究では、メタデータスキーマの構造的視点と意味的視点のそれぞれにおいてメタデータスキーマの変更履歴を追跡するための来歴記述モデルを提案し、かつ Semantic Web 環境に適した形式的記述で提案モデルを表すことにした。

以上のような考察に基づき、本研究では、DCAP の定義に基づきメタデータの構造的性質とメタデータ語彙の二つの視点からメタデータスキーマの来歴記述のモデルを定義すること、そして W3C が推奨する来歴記述のための標準 W3C PROV を基礎として定義することとした。W3C PROV は、Web 上の多様な資源への拡張性を持ち、加えて、PROV のデータモデル(PROV-DM)とオントロジー(PROV-O)が形式的に定義されているので、それらを基礎として来歴を RDF や OWL 等による形式的記述が行いやすいといった特徴がある。このことは、Web 上での来歴情報の流通そして保存にとって重要な意味を持つと考えられる。

本研究の早い段階で、筆者は PREMIS と PROV を組み合わせたメタデータスキーマの来歴記述の基本モデルについて検討した。そこから得た知見を基にして、本研究では W3C PROV をメタデータアプリケーションプロファイルとメタデータ語彙へ適用した。筆者は、メタデータの長期維持管理の視点から、メタデータアプリケーションプロファイルとして定義するメタデータの構造的制約とメタデータ語彙に含まれる個別の語句（メタデータターム）の意味の変更履歴の記述に焦点を当てた。PROV のデータモデルにおける来歴記述の重要な要素にエンティティ(Entity)とアクティビティ(Activity)がある。本研究では、メタデータアプリケーションプロファイルとメタデータ語彙の変更履歴を記述するためのエンティティとアクティビティを定義し、それをもとにデータモデルを定義することとし、最終的に、メタデータアプリケーションプロファイルにおける記述セットプロファイルの来歴記述とメタデータ語彙の来歴記述のために、それぞれ DSP-PROV および Vocab-PROV の二つのモデルを提案した。

DSP-PROV モデルは、記述セットプロファイルに定義した記述テンプレート(Description Template)、ステートメントテンプレート(Statement Template)、構造的制約(Structural Constraint)の変更履歴（修正・削除・追加）を記述する。したがって、DSP-PROV を用いることでそうした要素の変更をトレースすることができる。本研究では、開発した DSP-PROV モデルを米国デジタル公共図書館(Digital Public Library of America)のメタデータアプリケーションプロファイル(DPLA MAP)の定義文書の連続する 3 つのバージョンに適用した。なお、原定義文書は一定の形式で記述されたものであるが自然言語記述によって記述されたセミフォーマルなものである。この適用過程において、DSP-PROV による形式的記述を利用して原文書のバージョン間の矛盾を見つけることができた。こうしたことから、DPLA MAP の DSP-PROV による形式的来歴記述が、従来の自然言語による来歴記述に対して大きなメリットを持つと考えている。

本研究で提案したもう一つの来歴記述モデルである Vocab-PROV は、メタデータ語彙の変更履歴をシンプルな基本モデルの上に表現し、コンピュータ上で効率的に追跡できるようにすることを目的としている。 提案したモデルでは、メタデータタームとその定義記述の修正・削除・追加・代替を表現できる。Vocab-PROV は、DSP-PROV と同様、RDF を基礎として定義しているので、来歴記述の Web 環境での共有、SPARQL を使った検索やシンプルな推論などに利用することができる。

本研究では、開発した DSP-PROV と Vocab-PROV モデルに関して、それらの限界と実用性についても考察した。実際に、メタデータスキーマ上で起きる構造的変化と意味的変化は、この二つのモデルで示す基本的変化タイプよりも複雑になる可能性がある。しかし、本研究で提案したモデルではシンプルな変化を組合すことで多様な変化を表現することができることを前提にしており、メタデータスキーマの多様な変化にも対応できると考えている。そして、このモデルは、メタデータスキーマの来歴情報の追跡や長期維持の他、メタデータレジストリーの機能拡張、メタデータマッピング時に生じるエラーの発見等に利用できると考えている。

本研究から得た結論は以下のとおりである。(1)メタデータを人間と機械の両方が長期に渡って矛盾なく解釈できるようにすることはメタデータの長期利用の基本的要求である。メタデータの長期利用にはメタデータスキーマの長期利用性が求められ、それには

メタデータスキーマの長期の維持管理が必要である。(2)メタデータスキーマの長期維持のため、メタデータアプリケーションプロファイルとメタデータ語彙を長期間、矛盾なく維持管理することが、ネットワーク情報環境に依存する割合が益々高まっていくであろう将来に向けた重要な課題である。(3)Web 環境におけるメタデータスキーマの来歴情報の記述には、Web を介して異なるシステム間で来歴情報を相互運用できること、来歴記述を機械的に解釈し、追跡できることが要求される。(4)メタデータスキーマの構造的な変化と意味的な変化の分析は別々に行えねばならず、その一方、両者は同期して起きる可能性がある。(5)自然言語で記述されたセミフォーマルな来歴記述より、Web の標準に従った形式的な来歴記述は大きなメリットを持つ。

Table of Contents

List of Figures

List of Tables

1. Introduction

Libraries, archives, museums, data centers, government agencies, corporations, and individuals have been creating and managing a large number of collections of digital contents, which should be preserved for future use. It is widely recognized that digital objects should be kept usable over time and across communities. Researchers and practitioners are striving to make digital objects available and accessible to users over time. Digital preservation is crucial for keeping longevity of digital objects. Digital preservation is a cluster of many factors, which include financial, social, political, administrative and technological factors. What to be preserved is a basic question in the field of digital preservation. The diversity of digital objects exists among and within the types of digital objects. How long to preserve is another basic question and digital objects can be preserved for short-term, or middle-term or long-term according to their values. Digital preservation is "a game of probabilities". The preservation activities are undertaken to reduce or prevent the possibility of a preserved object from being lost or corrupted. However, there is no 100% guarantee that digital objects and their contents can be safely preserved in the long run (Wilson, 2017).

Digital preservation related issues have been discussed since 1990s. A variety of research projects, initiatives and efforts have been conducted to support longevity of digital objects. For instance, MetaArchive Cooperative as a digital preservation network for memory organizations, Open Archival Information System (OAIS) reference model (adopted as ISO standard 14721), Preservation Metadata: Implementation Strategies (PREMIS) metadata standard, CURL Exemplars in Digital Archives (CEDARS) project, Networked European Deposit Library (NEDLIB) project, Lots of Copies Keep Stuff Safe (LOCKSS) Program, Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP), Heritrix Web crawler project of Internet Archive, Australian Data Archive, UK data archive, Data Preservation Alliance for the Social Sciences (Data-PASS), and so forth. There are also many institutes for promoting research about digital preservation, such as International Internet Preservation Consortium (IIPC), Digital Curation Centre (DCC), Digital Preservation Coalition (DPC). Moreover, international conferences (e.g., International Conference on Preservation of Digital Objects, International Digital Curation Conference), journals on digital preservation (e.g., International Journal of Digital Curation, International Journal of Digital Libraries, Journal of Digital Information), and Web magazines (e.g., Ariadne, D-Lib Magazine) also promote research on longevity of digital objects. Those efforts have explored many research issues related to digital preservation. For example, sustainable digital preservation, preservation strategies (e.g., emulation, migration), preservation planning, risk

management, authenticity of preserved objects, data management, data lifecycle, schema evolution, file formats for long-term preservation, intellectual property rights, and so forth. These research issues help us understand digital preservation.

Digital preservation needs to deliver the past to the future in an authentic state. Digital preservation is not only about storage, backups, recovery, and access. It is not a one-shot effort and not an afterthought. There is a need to make continual efforts for on-going use of digital objects over time and adequate preparedness in advance. Digital objects can be preserved as a set of bit sequences. It is of importance to ensure that the bits remain intact over time. Unfortunately, the continued accessibility and usability of digital objects cannot be successfully guaranteed in the dynamic environments since digital objects are fragile. Even digital objects can be preserved in various mediums without damage or loss, it is difficult to make sure that users can interpret the contents of digital objects over time. Preserving bitstreams of digital objects alone is not sufficient for the long-term preservation of digital objects. There is a need to display digital objects in interpretable forms and keep the contents of digital objects interpretable by users (including both humans and machines) regardless of environments that may change over time. That is, simply preserving the bitstreams does not guarantee ongoing access to digital objects over time, which should be displayed in a meaningful form by the future users for long time. Digital preservation has been studied a lot to improve practices in the past decades. However, metadata preservation is still a new research area. It is necessary to understand why metadata preservation is required. To answer this question, the author will first give a brief introduction about the definition and roles of metadata.

What is metadata? Metadata (Greek: meta- + Latin: data "information") (Baca, 2008) is generally defined as "data about data". Metadata is "structured data about an object that supports functions associated with the designated object" (Greenberg, 2003). The traditional card catalogs and finding aids are metadata (Edward and Heather, 2014). And why does metadata matter? Metadata plays important roles in description, discovery, management and preservation of digital objects. The following paragraph explains why metadata of a preserved digital object should be preserved as well to keep the digital object alive for future use.

The OAIS reference model as an ISO standard (latest one is ISO 14721:2012) has broad applicability and serves as a framework for understanding the components and functions of an archive. It is widely accepted as an architecture of a long-term preservation system. Digital preservation needs metadata. The OAIS defines Preservation Description Information (PDI) as "the information that is necessary for adequate preservation of the content information and can be categorized as Provenance, Reference, Fixity, Context and Access Rights Information" (CCSDS,

2012). “These five kinds of information must be incorporated in digital preservation metadata” (Edward and Heather, 2014). PDI documents provenance and rights information of a digital object as administrative metadata. The white paper entitled “Preservation Metadata for Digital Objects: A Review of the State of the Art” and the report “Preservation Metadata and OAIS Information Model” with subtitle “A Metadata Framework to Support the Preservation of Digital Object” by the OCLC/RLG Working Group on preservation metadata address the importance of preservation metadata and review practices in the use of preservation metadata in the digital preservation community (OCLC/RLG, 2001; OCLC/RLG, 2002). The preservation of digital objects involves a variety of challenges, and metadata is one of them. Digital objects are preserved with their associated metadata. Metadata is one of the fundamental technologies that digital preservationists use to organize and retrieve contents in the digital preservation system. Metadata affects access to digital objects in the long term. It may be obvious that without metadata there is no access, since digital preservation systems will not be able to retrieve digital contents that are not described. Then, it will be difficult for users to find, identify, select and obtain digital objects. And without metadata in digital preservation process, context and authenticity of digital objects cannot be ensured either (Edward and Heather, 2014).

Given to the important roles and value of metadata, metadata longevity becomes a crucial issue. Metadata interoperability is still a big challenge in the research field of metadata longevity. Metadata longevity should keep metadata interoperable for data exchange among communities and across time. Keeping metadata interoperable over time, i.e., temporal interoperability of metadata, is the main concern of metadata longevity in this study. Temporal interoperability of metadata can be understood as active management of metadata to ensure ongoing access to and interpretability of metadata over time, with a purpose to communicate and connect metadata among past, current and future. There is a need to provide future users with appropriate information (e.g., contextual information, provenance information) to interpret metadata over time. Digital preservation related studies have built a firm foundation for metadata preservation. However, metadata preservation is not the same with digital preservation since metadata has its own features in the Web environment. It is required to make efforts to move the research field of metadata longevity forward.

Metadata should be consistently maintained after its manual or automatic creation in various forms in either conventional centralized or networked information environment. Metadata are maintained and stored in various databases such as relational databases, XML-based databases, Grid databases, and RDF stores. In the conventional environment, metadata are managed in closed systems. In the networked information environment, metadata is transferred as a digital object from a site to another and shared among those sites. To interpret these metadata, it is necessary to know

scheme of metadata (Rothenberg, 1998). A metadata schema is a (semi-) formal description scheme that defines syntactic, structural, and semantic features of metadata used for an application. A metadata schema defines implementation syntax and structural constraints of metadata, as well as metadata terms with their semantics from metadata vocabularies (Nagamori and Sugimoto, 2004). Dublin Core Metadata Initiative (DCMI) proposed Singapore Framework for Dublin Core Application Profile (DCAP) that is used as a generalized model of a metadata schema for an application and its related components, e.g., metadata vocabulary. The framework separates metadata terms and structural features of metadata. These components of a metadata schema should be consistently maintained across generations of technologies and users. This means, structural constraints including definitions of data structure, mandatory levels, and iteration constraints of description defined in a metadata schema should be consistently maintained. Moreover, definitions of terms and relationships between terms in a metadata vocabulary should be also consistently maintained.

Long-term maintenance of metadata schemas and metadata vocabularies requires to maintain their change history. As time passes, metadata schemas and metadata vocabularies are revised due to emergence of new requirements of resource description, development of technologies, and other reasons. Changes to metadata schemas and metadata vocabularies may cause inconsistencies in the future use of metadata. These changes include addition or deletion of a property, revision of value class, revision of mandatory level of a property, revision of meaning of a term, renaming of a term, revision of relationship between a term and another term, and so forth. These changes can be described in a provenance record by describing what an activity led to what kind of a change. In general, provenance (from French word “provenir”, “come forth, arise, originate”; from Latin “provenire”, “come forth, originate, appear, arise”) means origin or source or derivation of an object that can be work, data, etc. According to the definition of provenance given by W3C Provenance Working Group, provenance is a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. Provenance is used for many purposes, e.g., making judgments about information to determine whether to trust it, reproducing how something was generated (Moreau et al., 2013; Gil et al., 2013). Provenance of data provides proof of chain of data custody to ensure data authenticity. In the digital environment, provenance is necessary for preserving digital data since provenance assists in understanding the context of data, justifying trustworthiness of data and auditing inconsistencies or errors in data. Provenance of metadata describes how metadata came into being and its change history since its origination over time. Provenance of metadata provides contextual information about metadata, e.g., who created it, by what activity, for what purpose, and how it was organized

and processed over time. Provenance of metadata can be used to enrich contextual information of metadata. No provision of provenance may lead to loss of trust in metadata. Hence, it is important not only to provide but also to record the provenance information of metadata. There is necessity to clarify how to describe provenance, from where to start the provenance trail and how to maintain provenance. Therefore, metadata and its provenance are both critical to interpret specific metadata instances whose schema changes over time. Metadata along with its provenance needs to be consistently recorded and maintained over time for future use. However, provenance of metadata schemas is not well discussed yet in the digital presentation community.

The following two paragraphs briefly explain about the motivations of this study. In the Web environment, a wide range of W3C specifications (e.g., RDF, RDFS, OWL, SPARQL) and metadata frameworks (e.g., XML-based, RDF-based, and OWL-based metadata framework, such as Singapore Framework for Dublin Core Application Profile) have been developed to support metadata activities (Kashyap et al., 2008). Metadata is increasingly created and exchanged on the Web. Libraries, archives, and museums are also providing Web services to their digital collections using metadata. As mentioned above, a variety of research projects and initiatives in the digital preservation community have been complemented. It is evident from the relevant literature and practices that metadata has emerged as a vital part for the long-term maintenance of digital objects. However, previous research about metadata longevity is very limited. In the long run, there are many issues affecting access and use of metadata, such as economic issues, organizational issues, management issues, technology issues. The key issues affecting metadata longevity should be clarified. This is the first motivation to conduct this research with focus on management issues in metadata longevity. Based on the understanding of the state-of-the-art of research related to metadata maintenance, this dissertation mainly discusses metadata longevity from the following two aspects: long-term maintenance of metadata application profiles and long-term maintenance of metadata vocabularies.

As stated in OAIS and PREMIS, provenance is essential to authenticity of digital objects. Provenance description and provenance interchange have been discussed in library science, archival science, museum science, computer science, and so forth (Lemieux, 2016). Take the libraries and archives community as an example here. Library of Congress proposed “Explanation: DigProv (Digital Provenance) Extension Schema”, “DIGPROVMD: Digital Production and Provenance Metadata Extension Schema”, and “DigProv Data Dictionary: Audio-Visual Prototyping Project” to document provenance information. The archival standards such as General International Standard Archival Description (ISAD(G)), Encoded Archival Description (EAD), International Standard Archival Authority Record for Corporate Bodies, Persons and Families

(ISAAR(CPF)), and Encoded Archival Context (EAC) define the description elements for provenance information. There are already a wide range of models, ontologies, and vocabularies that can be used for provenance description, such as Open Provenance Model (OPM), Open Provenance Model Vocabulary (OPMV), Open Provenance Model OWL Ontology (OPMO), Open Provenance Model for Workflows (OPMW), Provenance Vocabulary (PRV), Vocabulary for Data and Dataset provenance (Voidp), Provenance, Authoring and Versioning Ontology (PAV), W7 Model, Provenir Ontology, BBC Provenance Ontology, W3C PROV standard, and others (Li and Sugimoto, 2014). However, existing technologies and standards are not specialized for metadata schemas. Specially, models for formal provenance description of metadata are not sufficiently explored. In the Web environment, there is a need to develop models for formal provenance description of metadata schemas in machine-readable and interoperable form, which can support automated and effective metadata maintenance. This is another motivation of this study. In the study, the author developed models for provenance description of metadata application profiles and metadata vocabularies, respectively. The models are proposed by applying W3C PROV standard to metadata application profiles and metadata vocabularies. The main reason for selecting W3C PROV standard is that W3C PROV is a Web-oriented provenance standard for provenance description and provenance interchange.

Inspired by the above two motivations, the author set the research goals to support metadata longevity as follows. One goal is to clarify key issues in metadata longevity. Another goal is to create models for formal provenance description of metadata schemas that enables machine-processable trace of revision history of metadata schemas. To achieve the first goal, the author analyzed features of metadata on the Web and risks affecting metadata longevity. As a result, the author identified long-term maintenance of metadata application profiles, long-term maintenance of metadata vocabularies, and risk management of metadata as the key issues in metadata longevity. To achieve the second goal, the author conducted provenance modeling for metadata schemas based on W3C PROV and DCAP. The author also applied the proposed provenance model for metadata application profiles to a case study. The specifics will be introduced in depth in the following chapters.

The main contents of this dissertation are structured as follows. Chapter 1 provides the background of the research. This chapter explains the necessity of formal provenance description of metadata schemas for metadata longevity. Chapter 2 states key issues in metadata longevity and provides meaning of basic concepts in the study. The research problem, goals and novelty of this research are also defined in this chapter. Chapter 3 introduces and reviews related literature. The author identifies the difference between related research and this research, and then indicates the

novelty of this study. Chapter 4 discusses digital provenance and metadata provenance. The author gives a primary idea to combination of metadata standard with provenance standard. Chapter 5 presents long-term maintenance of metadata schemas. The author developed DSP-PROV model to describe provenance of Description Set Profile with a basis on Dublin Core Application Profile. The author conducted a case study of Digital Public Library of America Metadata Application Profile to apply and evaluate the proposed DSP-PROV model. Chapter 6 addresses long-term maintenance of metadata vocabularies. This chapter classifies the primitive changes of metadata terms and their provenance description. Chapter 7 discusses lessons learned from this study and further research issues. The author presents the limitations and implications of the proposed models and other issues. The contents cover standardization and development of metadata application profiles, contextual metadata, provenance of research data, provenance of Linked Data, etc. Chapter 8 summarizes the main contributions and achievements of this study. The author also presents several suggestions on future work with open issues.

2. Long-term Maintenance of Metadata for Metadata Longevity

This chapter describes basic concepts used in this study. The author addresses research problems about temporal interoperability of metadata and explains the benefits of provenance description for metadata longevity. She clarifies the crucial issues in metadata longevity from perspectives of metadata application profile longevity, metadata vocabulary longevity and risk management in metadata longevity. She also points out the novelty of this research compared to previous studies.

2.1. Basic Concepts

2.1.1. Concepts Related to Singapore Framework for Dublin Core Application Profiles (DCAP)

Metadata is "(Structured) Data about Data". Metadata is extensively used as "description about anything" or simply "data". The ISO 15489-1 (information and documentation – records management) explains metadata as "data describing the context, content and structure of records and their management through time" (ISO 15489-1, 2016). **Meta-metadata** is metadata about metadata. For example, who created the metadata, when it was created, and how it was created are meta-metadata (Greenberg, 2003).

Singapore Framework for Dublin Core Application Profiles defines a set of descriptive components that are necessary or useful for documenting an application profile and describes how these documentary standards relate to standard domain models and Semantic Web foundations. Metadata practitioners had begun to experiment with the idea of Application Profiles since 1999. The Singapore Framework for DCAP defines the components of a metadata schema for an application and related components such as metadata vocabularies (Nilsson et al., 2008).

According to the Singapore Framework, a **Dublin Core Application Profile (DCAP)** is a packet of documentation that consists of Functional Requirements, Domain Model and Description Set Profile, Usage Guidelines and Encoding Syntax Guidelines. The first three components are mandatory and the last two components are optional. **Description Set Profile (DSP)** defines structural constraints of metadata (Nilsson et al., 2008). There are two levels of templates in a DSP. One is **Description Template (DT)** that contains the "statement templates that apply to a single kind of description as well as constraints on the described resource". Another one is **Statement Template (ST)** that contains "all the constraints on the property, value strings, vocabulary encoding schemes, etc. that apply to a single kind of statement" (Nilsson, 2008). **Structural Constraint (SC)**

defines structural features of metadata neutrally to any implementation syntax. Structural constraints include definition of data structure, mandatory levels, iteration constraints of description, and other constraints on properties and property values defined in statement templates. This study refers to the DSP itself and its components (i.e., DT, ST, SC) as a structural schema instance.

Metadata Schema is (semi-)formal description of a scheme which defines syntactic, structural and semantic features of metadata used for an application. Metadata schema is a typical meta-metadata (Li et al., 2015). Some metadata schemas are established as metadata standards by International Organization for Standardization (ISO) and National Information Standards Organization (NISO) due to their wide acceptance and usage, for instance, ISO 19115 standard for geographic data, NISO MIX XML schema for images. The term "metadata schema" is often used interchangeably with "metadata specification" and "metadata standard".

In general, a metadata schema uses metadata elements drawn from metadata vocabularies and establishes rules for the creation, use and management of metadata specifically regarding to the semantics, syntax, and optionality (obligation level) of values (ISO, 2016). It is generally understood to be a structured framework referring to data structures (Greenberg, 2005). In detail, a metadata schema can define the following points: (1) which elements are used to describe the resource, (2) if the elements are mandatory or optional, (3) if the elements are repeatable and how many times they can or must appear in a metadata description, (4) what is the value type or format of the elements, and other usage constraints. A metadata schema provides guidelines on the usage of the elements, identifies element obligations and other constraints, and provides comments and examples to assist in the understanding of the elements. The elements can be newly defined or extracted from one or more other existing vocabularies which may or may not be neutral to any application. For example, the Dublin Core metadata elements are defined neutral to any applications and used in many application profiles. This dissertation uses the word "element", which can be interchangeable with "property" and "attribute".

Metadata Vocabulary is set of metadata terms. It focuses on meaning of the terms and provides definitions of the terms and relationships between the terms (Patel, 2003), e.g., Dublin Core Metadata Element Set (DCMES) and Library of Congress Subject Headings (LCSH). This dissertation uses "metadata vocabulary" as a generic concept that includes two types, i.e., property vocabulary and value vocabulary. A **property vocabulary** is a set of terms expressing attributes of a resource and relationships between resources, which is often called metadata element set, e.g., DCMES and BIBFRAME vocabulary. A **value vocabulary** is a set of terms expressing classes of resources and encoding schemes of property values, e.g., Library of Congress Subject Headings

(LCSH). **Metadata Term** is controlled term defined for description of a metadata instance. There are two categories of metadata terms – property vocabulary terms and value vocabulary terms, which may be simply called property terms and value terms, respectively. A property term may be called an attribute or a descriptive element, e.g., terms defined in the DCMES. Every term included in a controlled vocabulary such as LCSH and MeSH, and the term that defines a class, type and encoding schemes of a property value is a value term.

In a broad sense, metadata schema includes **Metadata Application Profile** and **Metadata Vocabulary**. In a narrow sense, metadata schema can be viewed as same as metadata application profile. In the latter case, metadata vocabulary is separated from metadata schema and is used by metadata schema. A metadata application profile is oriented for an application, a community and a context.

2.1.2. Concepts Related to Metadata Longevity

Metadata Longevity is to keep metadata continuously accessible, usable, and interpretable for a long time by both humans and machines. It implies an active and continuous process, and concentrates on approaches with an emphasis on metadata interoperability across communities and over time. The longevity issues of metadata are mainly discussed from management perspective in this study.

Metadata Interoperability is the ability to exchange metadata without any special effort among different systems. There are many dimensions of interoperability, such as syntactic interoperability, structural interoperability, and semantic interoperability. This study focuses on temporal interoperability of metadata from the time dimension.

Temporal Interoperability of Metadata is interoperability of metadata over time. The following issues should be considered for temporal interoperability of metadata over its lifecycle: what happens to metadata since its origination, what causes the changes to metadata, how to keep metadata interpretable by both humans and machines regardless of changes in metadata application profiles and metadata vocabularies.

2.1.3. Concepts Related to Provenance

Provenance is defined by W3C Provenance Working Group as “a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. In particular, the provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it. In an open and inclusive environment such as

the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgements" (Moreau et al., 2013).

Digital Provenance is chronology or chronological information related to a digital object over its lifetime. Digital provenance typically describes agents responsible for the custody and stewardship of a digital object, key events that occur over the course of the digital object's life cycle, and other information associated with the digital object's creation, management, and preservation (PREMIS Editorial Committee, 2012).

Metadata Provenance is a record that typically describes responsible agents, influencing actions, associated events and other related information about metadata over its lifecycle (Li and Sugimoto, 2014). Both provenance of metadata application profiles and provenance of metadata vocabularies are metadata provenance.

Formal Provenance is provenance description in accordance with a well-structured scheme and recorded in an understandable and machine-processable form.

Semi-formal Provenance is provenance description following structured syntax and recorded in a natural language.

2.1.4. Concepts Related to W3C PROV

W3C PROV Standard is published by W3C Provenance Working Group and includes a set of documents, e.g., PROV-DM, PROV-O, PROV-CONSTRAINTS, etc. The standard refers to many aspects of provenance, such as modeling, serialization, exchange, access, validation, semantics, and reasoning (Moreau et al., 2015). In PROV, Entity and Activity are critical components to describe provenance. A **PROV Entity** is a physical, digital, conceptual, or other kinds of a thing. For instance, a Web page, a schema, or a vocabulary. A **PROV Activity** is something that occurs over a period of time and acts upon or with Entities. Activity is used to represent how an Entity comes into existence and how attributes of an Entity change to become a new Entity (Gil et al., 2013; Moreau et al., 2013). For example, publication of a paper, translation of a book, revision of a schema or a vocabulary.

2.1.5. Semantic Web Standards

Resource Description Framework (RDF) defines a model and syntax of metadata for World Wide Web. In the RDF data model, the basic unit of metadata is a statement expressed as a triple composed of <subject>, <predicate> and <object> (Schreiber and Raimond, 2014). An instance of metadata for an information resource is a set of triples where all the triples have the resource as its

<subject>. Property terms are used as a <predicate>. Value terms are used to specify class/type of a <subject> and <object> entities and used as an <object> (Li et al., 2015).

Web Ontology Language (OWL) is a knowledge representation language, designed to formulate, exchange and reason with knowledge about a domain of interest. OWL 2 ontologies provide classes, properties, individuals, and data values for modelling with OWL 2 (Hitzler et al., 2012).

Simple Knowledge Organization System (SKOS) provides a data model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading systems, taxonomies, folksonomies, and other similar types of controlled vocabulary. The SKOS vocabulary can be used to represent and publish concept schemes as machine-readable data on the Web (Isaac and Summers, 2009).

SPARQL Protocol and RDF Query Language (SPARQL) is a language and protocol for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware (Harris et al., 2013).

2.2. Temporal Interoperability of Metadata for Metadata Longevity

2.2.1. Metadata Transferred as a Digital Object on the Web

Metadata is a digital object in stored databases and interpreted by systems. Metadata in the networked information environment has features different from conventional metadata primarily designed for use in a single database or a set of databases. An instance of metadata on the Web is no longer an object enclosed in a database, but the instance is an object that is transferred from a site to another and shared among those sites. Metadata transferred as a digital object on the Web is a “first class object” and has features as follows. (1) Structural features: Metadata is typically structured according to a scheme. Structural features of metadata are assertions about data structure, mandatory levels, iteration constraints of description, and so forth. Such assertions represent attributes and values of resources in machine-readable form. (2) Syntactic features: Metadata can be serialized in different syntaxes, e.g., HTML, XML, RDF/XML, Turtle, JSON, and JSON-LD (Greenberg, 2003). (3) Semantic features: The elementary semantics of metadata are specified and defined in a metadata vocabulary. The meaning of every metadata term and the relationships between terms are identified as the semantic features of metadata. URI is used as the base scheme to identify a term in the Linked Open Data (LOD) environment.

Digital resource both in a database and on the Web, may be unidentifiable and irretrievable without metadata. Metadata exchanged and transferred as a digital object is at risk of being unusable

in the networked environment because of dynamic factors, such as unstable identifiers. Therefore, in the long run, metadata should be continually managed to ensure its availability, quality, persistence and permanence over time. However, metadata longevity is quite difficult to guarantee over time and there are still no well-established approaches for metadata longevity. The content, semantic, structure, and provenance of metadata should be maintained for metadata longevity.

2.2.2. Metadata Interoperability and Temporal Interoperability of Metadata

The "Interoperability levels for Dublin Core Metadata" has been identified as the following four levels of metadata interoperability. "At level 1, applications use data components with shared natural language definitions. At level 2, data is based on the formal semantic model of the W3C Resource Description Framework. At level 3, data is structured as Description Set (records). At level 4, data content is subject to a shared set of constraints (described in a Description Set Profile)" (Nilsson et al., 2008). According to Dublin Core metadata interoperability, using standardized metadata application profiles and metadata vocabularies are recommended for achieving better structural interoperability and semantic interoperability of metadata.

This study deals with issues related to metadata longevity, which focuses on keeping metadata interpretable by humans and machines over time. Metadata application profiles define data structure and metadata constraints while metadata vocabularies define semantics of metadata. Long-term maintenance of metadata application profiles and metadata vocabularies is required for consistent maintenance of structural and semantic features of metadata. Metadata application profiles, metadata vocabularies and metadata records may be changed over time. The dynamic environments bring changes to them and the changes should be also properly recorded for consistent maintenance of metadata over time. And recording their change history as provenance descriptions is beneficial to long-term maintenance of metadata.

The research problem addressed in this study is that there are no well-developed models for metadata provenance to support the longevity of metadata. The author attempted to propose models for formal provenance description of metadata application profiles and metadata vocabularies, which describes change history of metadata application profiles and metadata vocabularies for metadata longevity, respectively. This study is aimed to develop provenance models for metadata longevity by applying W3C PROV standard to metadata application profiles and metadata vocabularies. This study provides answers to the following research questions: What is the requirement of metadata longevity? Why need to keep metadata interpretable over time? Why metadata application profiles and metadata vocabularies should be consistently maintained? What kind of risks affect metadata longevity? What are the primitive changes to metadata application

profiles and metadata vocabularies? How to formally record these changes as provenance description? How to formally describe provenance of metadata application profiles and metadata vocabularies in machine-processable form by using Web standards?

2.2.3. Why Formal Provenance Description for Metadata Longevity

Provenance has been studied in different domains. In the archival and museum community, provenance is widely used for denoting ownership. In archival systems, provenance is adopted to ensure data trustworthiness. The research interest in provenance has been increasing and many working groups related to provenance have been established. The W3C Provenance Incubator Group (2005-2010) and W3C Provenance Working Group (2011-2013) have made a lot of efforts in developing standard for provenance representation. The International Provenance and Annotation Workshop (IPAW) is a biannual workshop since 2006 and concerns about data provenance, data derivation, and data annotation. The workshop on Theory and Practice of Provenance (TAPP) also facilitates the development of provenance research. DCMI Metadata Provenance Task Group implemented "Dublin Core to PROV Mapping" and published it as a W3C working group note. Data Observation Network for Earth (DataONE) project to support discovery of earth data and environmental data established Scientific Workflows and Provenance Working Group and made effort to develop provenance management architecture for scientific data processing systems. Provenance of research data is also addressed by researchers to facilitate data reproducibility. The Research Data Provenance Interest Group on Research Data Alliance focuses on comparison and evaluation of models for data provenance.

It is already recognized that provenance is crucial to longevity of digital objects according to the OAIS reference model and PREMIS metadata standard in digital preservation community. Provenance is widely used for data trust judgement, data quality assessment, data error checking, data reproducibility, revelation of Web pages' revision history, and so forth. W3C Provenance Incubator Group reported use cases of provenance, which refer to three dimensions of provenance proposed by this group, i.e., content, management and use (W3C Use Case Report, n.d.). W3C Recommendation titled "Data on the Web Best Practices" (Lóscio et al., 2017) recommends providing complete information about the origin of the data and changes history, and explains the reason for providing data provenance information. Understanding the origin and history of data helps determining whether to trust data and provides important interpretive context.

The Semantic Web is designed to represent information in a machine-readable format. The machine-readable data provenance can be provided using an ontology recommended to describe provenance information, such as W3C's provenance ontology (Lóscio et al., 2017). The use of

Semantic Web technologies has been advocated to facilitate provenance acquisition, representation, and reasoning. The triple structure of RDF simplifies graph representation. Many researchers advocate the use of RDF to represent provenance information (Moreau, 2010). In the Semantic Web environment, the ability of processing and exchanging provenance among different systems is required. The advantage of the formal description of a metadata schema over conventional change-logs is automated auditing to help find errors and inconsistencies between the versions of the metadata schema. Hence, Semantic Web standards PROV and RDF are used as bases to formally describe metadata provenance, which is machine-processable for provenance interchange on the Web.

2.3. Management Issues in Metadata Longevity

Metadata is created according to its metadata schema, which often uses terms from metadata vocabularies to describe a resource. Metadata standard and terms are beneficial for understanding metadata. Standards of metadata and dictionaries of terms used in metadata should be stored to guarantee usability of digital data (content and metadata) after long time. “Without them after long time information contained in metadata might be difficult to understand” (Traczyk et al., 2017). Without metadata schema and metadata vocabulary, users may misunderstand data structure and meaning that metadata holds. A metadata schema should be preserved as well as metadata instances created according to the schema, and the adopted metadata vocabulary should be also preserved to keep semantics of metadata interoperable.

The author tried to determine the crucial management issues affecting metadata longevity. This section is devoted to discussions on metadata longevity from the perspectives of long-term maintenance of metadata application profiles and metadata vocabularies as well as potential risks in their long-term maintenance.

2.3.1. Long-term Maintenance of Metadata Application Profiles

Metadata application profiles have been developed in a wide range of domains for various purposes, for instance, DataCite metadata schema for resource citation and retrieval purposes, Asset Description Metadata Schema for describing assets (the narrower meaning of metadata schema is used here, and it does not distinguish metadata application profile and metadata schema). An application profile defines rules to describe a resource. Application profile supports resource description, metadata creation, metadata interoperability and metadata sharing. An application profile provides a guidance for metadata creation for a specific domain or a type of resource. Every

element defined in an application profile for a community is exchanged with the same meaning. Therefore, a well-defined and interoperable application profile can facilitate interoperability and sharing of metadata among systems.

An application profile is usually defined in a document that can be created in various forms, e.g., CSV, XML, RDF/XML. Application profiles are preserved as a document for human readers in the conventional maintenance environment of metadata. In the state-of-the-art Web environment, metadata application profiles are digital objects transferred over networks and are no longer simple document-like objects. With the development of Web standards, metadata application profile description has been transiting from semi-formal description in a natural language to formal description in a machine-processable language. The Web standards (e.g., RDF and OWL) assist in metadata application profiles description in a machine-processable form, which brings new requirements of long-term maintenance of metadata. The effective and consistent maintenance of metadata application profiles is needed in networked information environment. The constraints (for example, mandatory levels, iteration constraints, usage constraints) defined in a metadata application profile should be consistently maintained.

Metadata application profile can be developed by subject discipline and technical professionals in data representation and data processing. A metadata application profile is designed to meet community needs, which may change due to many reasons, such as emergence of advanced technologies, changes of resource scope, changes of resource description requirements. Different versions of a metadata application profile will be released if there are major changes made to the metadata application profile and the changes are approved based on the community consensus. Once a metadata application profile is published, its maintainer should continually maintain the application profile. The changes to metadata application profiles should be properly recorded to prevent inconsistencies in the future use.

Take the DataCite metadata schema as an example. The schema is “a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions”. The latest version of this schema is Version 4.0 released in 2016. It has previous versions, e.g., Version 3.1 released in 2014, Version 3.0 released in 2013, Version 2.2 and 2.1 and 2.0 released in 2011. There are changes between its two consecutive versions and these changes are recorded in a natural language. The following are some change examples from Version 3.1 to Version 4.0: “Changing resourceTypeGeneral from optional to mandatory”; “Addition of new optional subproperties for creatorName and contributorName: familyName and givenName” (DataCite Schema, n.d.). In the Web environment, there is a need to

record changes to metadata application profiles in machine-processable form for effective auditing errors or finding inconsistencies in metadata.

2.3.2. Long-term Maintenance of Metadata Vocabularies

Using a metadata vocabulary is beneficial to searching, finding and sharing metadata of resources. Memory institutions (including libraries, archives, museums) use controlled vocabularies, thesaurus, classification schemes, and name authorities for resource description, bibliographic organization and bibliographic control. For example, LCSH used in library community, Getty Art and Architecture Thesaurus (AAT) used in museum community. Some widely-used metadata vocabularies are identified as standards. For instance, DCMES has been approved as ISO standard 15836 and ANSI/NISO standard Z39.85-2012, which is widely used for metadata description of resources to exchange information in the networked environment.

Metadata vocabularies need long-term maintenance for future use. When a newly defined version of a metadata vocabulary is published, usually there are some changes from its previous version, e.g., renaming of a term, addition or deletion of a term. In addition, the meaning of a term may be changed, relationship between terms may be revised, a composite term may be split to single terms, or a set of single terms may be merged into one composite term, a bibliographic reference cited in a usage comment may be updated, the status assigned to a term may be changed. Take the term addition in DCMES vocabulary as an example. The standardized DCMES is composed of fifteen core elements. The changes made to DCMES elements between 2001 and 2006 were maintained by the DCMI Usage Board in light of the DCMI Namespace Policy. Each decision of DCMI Usage Board is assigned a URI, and links are created to support documentation, decision texts, and to the historical term declarations of any metadata terms affected by the decisions (Baker, 2007; Baker, 2004).

Clarifying requirements and issues in maintaining metadata vocabularies is useful for metadata maintenance. Stability is a key concept for long-term maintenance of metadata especially for metadata vocabularies published on the Web. The stability of the vocabulary URI and term URI is essential on the Web. The Web environment brings other issues for the long-term maintenance of metadata vocabularies that are represented in RDF by using Web standards such as RDF schema (RDFS), OWL and SKOS. Maintenance of term identifiers and their stability is an important issue referring to access of metadata terms on the Web.

Well-defined and sustainable metadata vocabularies can lead to better interoperability and harmonization across institutions and over time. Therefore, maintainers of vocabularies should make sustainable policies referring to namespace policy, publication policy, change policy, and so

forth. Baker discussed a set of requirements for vocabulary preservation and vocabulary governance. Metadata vocabulary creators and maintainers are recommended to reuse existing and well-known metadata vocabularies to improve semantic interoperability of metadata. Vocabulary managers, standard bodies, and memory institutions can work together for global governance (Baker et al., 2013).

Not only the documentations of metadata vocabularies themselves but also the changes made to them are worthy of being recorded. Proper change documentation should include sufficient meta-information to assist users in understanding the change, the requirements driving it, and its potential consequences (Baker and Alistair, n.d.). Provenance description of metadata vocabularies describing change history of metadata vocabularies is a kind of meta-information that can provide contextual information for these changes, e.g., what had been changed, how it was changed.

2.3.3. Risk Management in Metadata Longevity

Handing risks is a significant task for long-term preservation of digital objects. Risk management is required to ensure continual monitoring of potential risks and minimize their possible effects. In the OAIS reference model, risk management is an essential part of preservation planning (Hein and Schmitt, 2013). The Simple Property-Oriented Threat (SPOT) and Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) have been proposed as risk assessment methodologies. The SPOT model focuses on safeguarding against threats to six essential properties (i.e., availability, identity, persistence, renderability, understandability, and authenticity) of digital objects. The SPOT provides a simple model for risk assessment on these six properties. DRAMBORA is a toolkit for a digital repository audit including a list of over 80 examples of potential risks to digital repositories (Dappert, 2016; Vermaaten et al., 2012; DPC, n.d.). DRAMBORA provides a risk-based approach to enable repositories to monitor how they are handing the risks associated with preservation. SPOT and DRAMBORA are mainly used for risk management in digital preservation community. However, SPOT and DRAMBORA do not classify risks in metadata longevity.

Risks in metadata longevity actually exist, and these risks can be technological, physical, organizational, legal, financial, political, etc. These risks should be managed to mitigate the likelihood of their occurrence. Risk management for keeping metadata and its schema safe is a crucial research issue. Without managing potential risks, problems in reusing metadata might be caused. Loss of metadata schema (metadata schema including metadata application profile and metadata vocabulary in a broad sense is used here) and no provision of its provenance information may result in loss of data meaning, difficulty in identifying data authenticity, inability of data reuse,

and cost of recreation and recovery of data. Therefore, metadata longevity requires preventing metadata from potential risks in long-term maintenance of metadata. Metadata should be kept interpretable for the future and risks in the longevity of metadata application profile and metadata vocabulary should be detected. The author gives risk analysis with emphasis on risks in long-term maintenance of metadata application profiles and metadata vocabularies by following the steps identified by general risk management standard ISO 31000:2009.

ISO 31000:2009 provides a guideline for managing risks, which can also guide risk management in metadata longevity. Risk is defined as the combination of the probability of an event and its consequences. According to ISO 31000:2009, risk assessment comprises three steps including risk identification, risk analysis and risk evaluation. The step of risk identification is to identify sources of risk, areas of impacts, events and their potential consequences. Significant causes and factors that have impact on metadata longevity as well as consequences should be considered in the step of risk analysis. Risk evaluation step determines treatment to these risks (Leitch, 2009; ISO, 2009). According to the steps defined in this standard, the author generally identified risks in longevity of metadata schemas. (1) Metadata schema describing a resource may be unknown, improperly recorded, lost, changed, or obsolete. (2) Metadata schema describing a resource may be improperly maintained and their revision history may not be consistently recorded. (3) Provenance information about the resource and its metadata schema may not be consistently recorded in machine-processable form. (4) Resource identifier may be inconsistent or instable.

The author proposes strategies to avoid the risks in metadata longevity as follows: (1) preserving the documents of metadata schema, (2) recording and maintaining metadata schema along with their revision history, (3) recording provenance of metadata schema, and (4) creating sustainable identifiers schemes. Table 2.1 provides a brief analysis to these risks referring to their causes, consequences and treatments.

Table 2.1: Risks in longevity of metadata.

Risk 1 Metadata schema may be unknown	Cause	no recording of the name of metadata schema
	Consequence	user cannot find definitions of data structure and terms
	Treatment	recording the used metadata schema
Risk 2 Metadata schema may be improperly recorded	Cause	incorrect recording
	Consequence	misunderstanding of metadata; inconsistency in metadata mapping
	Treatment	error checking
Risk 3 Metadata schema may be lost	Cause	lack in preservation of metadata schema; failure in preservation of metadata schema
	Consequence	user cannot understand metadata
	Treatment	successful preservation of metadata schema
Risk 4 Metadata schema may be changed	Cause	new versions of metadata schema are released
	Consequence	metadata records are not kept invalidated
	Treatment	consistent recording of occurred changes
Risk 5 Metadata schema may be obsolete	Cause	failure transformation between formats
	Consequence	the contents in metadata schema are lost
	Treatment	file format migration; using tools for format validation
Risk 6 Metadata schema may be improperly maintained	Cause	human operation mistake
	Consequence	loss of metadata schema
	Treatment	conducting error auditing
Risk 7 The changes to metadata schema are inconsistently recorded	Cause	not all changes are recorded
	Consequence	maintainer cannot continually track the chain of the revision history of metadata schema
	Treatment	changes detection, recording and tracking
Risk 8 Identifiers used in metadata schema may be instable	Cause	identifier scheme is changed
	Consequence	metadata cannot be accessed
	Treatment	using persistent identifiers

2.4. Research Goals and Research Challenges

The overall goal of this study is to facilitate metadata longevity through long-term maintenance of metadata application profiles and metadata vocabularies. As mentioned above, the author clarified management issues of metadata longevity and requirement of metadata longevity

from time dimension, with a purpose to propose models for provenance description of metadata schemas to support long-term maintenance of metadata.

It is a big challenge to keep metadata interpretable by humans and machines over time. There are difficulties and problems in keeping contents and semantics of metadata interpretable in the dynamic information environment and over time. For example, loss of metadata schemas and no provision of their provenance may lead to difficulty in understanding metadata; inconsistencies caused by changes to metadata schemas over time. The change history of metadata schemas, including changes in structural constraints and semantics, should be consistently recorded as provenance descriptions for long-term maintenance of metadata schemas. Therefore, the author is attempting to keep change history of data structure and meaning of metadata readable and understandable both by humans and machines through formal provenance description of metadata, regardless of changes to metadata application profiles and metadata vocabularies over time. Another research challenge lies in how to generalize model for provenance description of metadata schemas and how to consistently maintain their revision history. This study proposes models with basis on the features of metadata application profiles and metadata vocabularies aligned with their primitive changes.

2.5. Research Novelty

Researchers have paid a lot of attention to longevity of digital objects. Metadata preservation is still a new issue which is different from preservation of those digital objects. Based on the survey results in this study, the author has found that less studies related to metadata preservation have been carried out. Literature review in detail will be given in Chapter 3. Metadata curation can be broadly interpreted as active maintenance of metadata and appraisal of metadata for both current and future use over its entire life cycle. Metadata curation involves maintaining, preserving and adding value to metadata throughout its lifecycle (DCC, n.d.). And metadata management is the sum of activities designed to create, preserve, describe, maintain access, and manipulate metadata (Westbrooks, 2005). Though there are studies related to metadata curation and metadata management (Mayernik, 2016; Shaon and Andrew, 2008; Sun microsystems, 2005; Shaon, 2005), their perspectives (e.g., lifecycle management, metadata quality) are quite different from this study. This study is carried out from the perspective of temporal interoperability of metadata and the view of metadata provenance description to discuss metadata longevity. In practice, metadata repository can assist in metadata collection and metadata storage. Metadata registry can provide function of maintenance of metadata application profiles and metadata vocabularies. However, metadata repository and metadata registry do not ensure metadata longevity and temporal interoperability of

metadata. As introduced in Section 2.3, several management issues related to metadata longevity are raised up. There is necessity to explore solutions to these management issues. Given to important roles of provenance in the longevity of digital objects, the author proposes to formally record provenance description of metadata for metadata longevity over time.

Provenance describes a series of events and activities happened on a digital object and is required for data trustworthiness in digital archival systems. However, provenance description for metadata longevity is not well discussed in the community of digital preservation. In this study, W3C PROV is selected for provenance description of metadata since that PROV is developed for provenance description and provenance interchange on the Web. It is already recognized that PROV can be applied to specific applications or domains due to its extendibility. Although W3C PROV standard has been applied to describe various kinds of data provenance (e.g., provenance of workflow, research data, and climate data), there are still no models for provenance description of metadata application profiles and metadata vocabularies. Therefore, another novelty of this study lies in the proposed models for formal provenance description of metadata to facilitate long-term maintenance of metadata schemas. The proposed models are novel, and they enable trace of primitive changes of metadata application profiles and metadata vocabularies.

3. Literature Review

This chapter reviews the relevant research. It covers previous literature on metadata curation, metadata management, and metadata interoperability. Research related to provenance has mainly involved the provenance models, provenance tracking, provenance uses in libraries, archives and museums, and provenance issues on the Semantic Web.

3.1. Metadata Curation and Metadata Management

There is strong emphasis on digital curation, data curation, and digital preservation in the research community (Poole, 2016). Metadata has been recognized as the key function of curation and preservation. Metadata curation "may be defined as an inherent part of a digital curation process for the continuous management (which involves creation and/or capturing as well as assuring overall integrity of metadata amongst other things) and preservation of metadata records over their life cycles" (Shaon, 2008). Shaon (2008) proposed a metadata curation model embedded in the OAIS reference model with functions of metadata ingest entity, metadata quality assurance entity, the metadata versioning entity, and metadata management entity. Mayernik (2015) outlined five categories of institutional carriers to analyze how data management, curation, and preservation practices emerge, evolve, and transfer within and across scientific institutions. Data practices and curation vocabulary (DPCVocab) consisting terms about research data practices, data and curation in earth and life sciences has been developed. DPCVocab provides a common vocabulary for interactions among curators, data producers, system developers, and other stakeholders in the curation process (Chao et al., 2015). However, the approach is a conceptual solution and needs test in a digital curation system. There are still no comprehensive and effective approaches to metadata curation.

Ball (2012) comprehensively reviewed main lifecycle models for data management including DCC curation lifecycle model, I2S2 idealized scientific research activity lifecycle model, DDI combined life cycle model, ANDS data sharing verbs, DataONE data lifecycle, UK data archive data lifecycle, Research360 institutional research lifecycle, and capability maturity model for scientific data management. Metadata management refers to the activities associated with ensuring the proper creation, storage, and control of metadata (metadata management white paper, 2005). Sen (2004) summarized the history of metadata management from file systems since 1960s to creation of metadata warehouse after 2000. Kim (2005) presented the difficulties in metadata management referring to metadata definition and management, technology and standards. He also listed up the basic set of facilities in a metadata management system.

The application and development of Web technologies (e.g., emergence of LOD and Semantic Web) bring new challenges for metadata curation and metadata management. The long-term usability of LOD is an emerging issue. LOD are machine-readable following Web standards and protocols, such as RDF, SPARQL. LOD are digital-born objects and structured data that change over time. Their dynamic characteristics bring the persistence issue.

Auer et al. (2012) stated the challenges of preserving LOD including provenance problems relating to the evolution of LOD datasets. They presented requirements of management of temporal and provenance annotations for constant accessibility of LOD. They proposed a distributed and service-based infrastructure for LOD preservation, which includes change detection, provenance support and other functionality. Papastefanatos (2013) presented LOD preservation and long-term accessibility issue, and proposed a framework integrating provenance tracking, change detection and quality control for management of LOD evolution.

The EU-funded "Preserving Linked Data" (PRELIDA) project started in 2013 for two year's research on Linked Data preservation. The project reports identify differences and analyze the gap between Linked Data preservation and digital preservation. They pointed out OAIS "does not ensure consistency or interoperability between implementations" and presented challenge to preservation of Linked Data (Giaretta et al., 2014; Grigoris et al., 2014). The project also gained insights and issues related to long-term usability of Linked Data, for instance, change management, data evolution.

Memento protocol is specified in RFC 7089 and defines interoperability for access to resource versions based on a resource's generic URI as it existed at a specific moment in time. The protocol is used to deal with archiving of different versions of Web resources (Auer et al., 2012). It has been adopted by many major publicly accessible Web archives, for example, Memento compliant DBpedia archive.

Researchers paid a lot of attention to data curation and data management in the past few years. They discussed the issues from the perspectives of lifecycle and data quality. From existing practices, the author learned that building data/metadata repositories, digital archives, and data/metadata management systems are options to manage and store data/metadata. Depending on operational and practical requirements, metadata can be embedded with the data, or stored separately from the data in a classic relational database or in an RDF triple store. Research issues about metadata curation are usually not treated as a separate issue, instead metadata roles are addressed within general data curation, research data curation, and data curation in specific domains.

The effective management of metadata is critical to data lifecycle management. Although metadata curation and metadata management have received attention from researches, keeping the

digital content of metadata persistent and interpretable is still a difficult problem to solve in practice. Methods and technologies for effective management and safe preservation of metadata for long time should be developed both for closed system and open Web environment.

3.2. Maintenance of Metadata Vocabularies

Semantic Web Best Practices and Deployment Working Group Vocabulary Management Task Force established in 2004 at the World Wide Web Consortium (W3C) has developed best practice guidelines and principles for publishing RDF vocabularies on the Web. Their achievements contribute a lot to the development and maintenance of vocabularies especially RDF vocabularies in the Web and Semantic Web environment (Kendall et al., 2008). For instance, identifying metadata terms using URIs, identifying the historical version of a vocabulary or its terms (e.g., provenance documentation), declaration of terms using a formal and machine-processable schema language.

The DCMI Vocabulary Management Community started the special session with the theme of vocabulary management at the international conference on Dublin Core and Metadata Applications (DC-2011). And later, DC-2013 continued this work and held another special session on vocabulary management. These sessions discussed crucial issues for maintenance of metadata vocabulary, such as persistent URLs, namespace policy, publication policy, tracing of vocabulary history, vocabulary preservation. The community addresses creation, maintenance, versioning and sharing of vocabularies and provides guidance to metadata practices, which has identified a range of management issues to be considered. Furthermore, a set of requirements for vocabulary preservation and governance have been presented: each term in a vocabulary is cited by a URI and resolvable to a formal, machine-readable representation of the term meaning; policies related to maintenance, copyright, and versioning are made available; reuse of the existing vocabularies; cooperation between memory institutions and vocabulary maintainers (Baker et al., 2013).

Linked Open Vocabularies (LOV)[1] as part of the DataLift project was launched and hosted by the Open Knowledge Foundation since 2011. The LOV initiative plays vital role in the vocabulary ecosystem. LOV gathers and provides the information such as interconnection between vocabularies, versioning history and maintenance policy (Vandenbussche et al., 2017). Kunze et al. (2017) presented their work about development of a persistence vocabulary and solutions for identifier technology for the objects that the scientists want to reuse for the long-term.

[1] Please see http://lov.okfn.org/dataset/lov

As stated in the previous research activities, the requirements for maintenance of metadata vocabularies in the Web environment reached to a common consensus. Stability and persistence are the main concern for long-term maintenance of metadata vocabularies. Despite the lack of change history and provenance may limit the reuse of metadata vocabularies, metadata terms in metadata vocabularies are still often treated and examined from a static rather a dynamic perspective in practice. Therefore, metadata vocabularies need proper documentation of their changes and contextual information with provenance to assist users in understanding the changes and the development of metadata vocabularies.

3.3. Metadata Registries for Metadata Interoperability

The standards for metadata registry have been well developed, for example, the International Standards Organization/International Electrotechnical Commission 11179 (ISO/IEC 11179), metadata registries standard developed by ISO/IEC JTC1 SC32 WG2 Development/Maintenance. ISO/IEC 11179 has been a vital standard for the development of metadata schemes for digital resources. The standard discusses and introduces fundamental ideas of data elements, value domains, data element concepts, conceptual domains, classification schemes; provides guidance on how to develop ambiguous data definitions. Other standards guiding metadata scheme development include ISO/IEC 20943, Procedures for Achieving Metadata Registry Content Consistency; ISO/IEC 20944, Metadata Registry Interoperability and Bindings; ISO/IEC 18038, Identification and Mapping of Various Categories of Jurisdictional Domains. These standards provide guidance for the development of metadata registries that are crucial for metadata interoperability. There is still no well-recognized standard for the longevity of metadata schemas, although there are de-facto and international standards designed for interoperable metadata such as Dublin Core Application Profiles and standards for metadata registries.

Metadata registries manage, store and provide search and/or browse services for the registered definitions of metadata vocabularies and metadata application profiles. Metadata registries play crucial roles in the management and sharing of metadata terms, metadata vocabularies and metadata application profiles across communities and over time (Dunsire, 2012). The metadata community has made achievements in the development of metadata registries, such as CORES registry, MetaBridge registry,[2] DCMI metadata registry,[3] Open Metadata Registry (OMR)[4], Resource

[2] Please see https://www.metabridge.jp/infolib/metabridge/menu/?lang=en
[3] Please see http://dcmi.kc.tsukuba.ac.jp/dcregistry/
[4] Please see http://dcmi.kc.tsukuba.ac.jp/dcregistry/

Description and Access (RDA) registry.[5] The reuse of existing metadata terms is essential to improve metadata interoperability. Although metadata interoperability is an important aspect for long-term maintenance of metadata, metadata registries do not ensure the long-term use of metadata that covers many aspects. Management aspects, economic aspects, organizational aspects, and technological aspects all have impact on metadata longevity. This study focuses on the long-term maintenance of metadata application profiles and metadata vocabularies.

Management and use of provenance information of metadata vocabularies and metadata application profiles have not been well discussed except issues related to versioning control. OMR provides service to vocabulary owners and managers about the versioning and change tracking of their registered vocabularies. The information about changed time, action, and the vocabulary maintainer who made the change are accessible on OMR history page. RDA vocabularies (element sets and value vocabularies) are maintained in the RDA Registry based on OMR with a combination of Git and GitHub. RDA Registry supports the semantic versioning of RDA vocabularies. The version designations follow the general principles of semantic versioning. GitHub provides the changes list of released RDA vocabularies in natural language, e.g., lists of "Adds new RDA entities", "Adds new RDA elements", "Adds new constrained RDA elements", "Deprecates published RDA elements", "Adds value vocabularies" and "Renames value vocabularies" (Phipps et al., 2015). However, these changes of RDA vocabularies are not kept interpretable to machines over time.

In the long run, a metadata schema for an application along with used vocabularies evolves and is exchanged for communication with future users. Changes in metadata schema and metadata vocabulary may cause inconsistencies in the long-term use of metadata. The consistent maintenance and change tracking of the structural constraints of metadata and semantic definitions of metadata are both required for metadata longevity.

3.4. Perspectives of Provenance

Provenance has gained a lot of attention as summarized in Table 3.1, which shows provenance related research in diverse areas, such as archival science, library and information science, computer science, cognitive science, and others (Lemieux, 2016).

In the archival science, General International Standard Archival Description (ISAD(G)), Encoded Archival Description (EAD), International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR(CPF)), and Encoded Archival Context (EAC)

[5] Please see http://www.rdaregistry.info/

cover provenance for arrangement of archival materials. The International Research into the Preservation of Authentic Records in Electronic Systems (InterPARES) project addresses the importance of provenance for keeping trustworthiness of digital records (Niu, 2013).

Provenance is used to identify authorship of works and origin of resources in the museum community. The Getty Provenance Index Databases[6] provide search services for provenance of archival inventories, auction catalogs and dealer stock books. The CIDOC Conceptual Reference Model (CRM) in the museum community has also been extended to model provenance information of digital objects (Theodoridou et al., 2010), for instance, CRMdig[7] model for provenance metadata.

In library and information science, provenance of rare books and research data have been discussed. For instance, the Council of European Research Libraries (CERL)[8] website provides search services of owners of old books. Provenance of research data is crucial for data reproducibility as a service at research libraries. The Research Data Provenance Interest Group[9] on Research Data Alliance launched in 2014 focuses on tracking provenance for research data, such as comparison and evaluation of models for data provenance, maintenance of identity through the data lifecycle.

In computer science, data provenance is a key issue especially in workflow and databases for data transparency, data quality and computational reproducibility (Simmhan et al., 2005). In geoscience, provenance description of climate change data and geographic data has been discussed, where W3C PROV is used as their base model (Masó et al., 2015; Tilmes et al., 2013). Provenance can be used to build trust in goods and supply chain in business community. For example, PROVENANCE[10] platform provides trace of digital history of registered products. Provenance associated with social media statements can be used to dispel rumors, clarify opinions, and confirm facts (Barbier et al., 2013).

In the evolving and dynamic metadata ecosystem, it is necessary to know how a metadata schema is derived from its origination to a particular version. This study is aimed to define a formal provenance description scheme for structural features of metadata schemas called application profiles. It focuses on description of change history of application profiles as provenance data exchangeable in the LOD environment.

---

[6] Please see http://www.getty.edu/research/tools/provenance/search.html
[7] Please see http://www.cidoc-crm.org/crmdig/home-2
[8] Please see http://www.cerl.org/resources/provenance/main
[9] Please see https://www.rd-alliance.org/groups/research-data-provenance.html
[10] Please see http://www.provenance.org/

Table 3.1: Overview of provenance research.

Domains/Fields/Communities	Research Focus
Digital Preservation	Authenticity of digital objects
Archival Science	Arrangement of archival records
Museum Science	Ownership of arts
Library and Information Science	Authorship of rare books; Provenance of research data
Computer Science	Data transparency, data quality, data reproducibility; provenance of Linked Data; provenance in workflow and databases
Cognitive Science	Visual analytics
Geoscience	Provenance of climate change data and geographic data
Business	Provenance of products to provide trust in goods
Social Media	Provenance of social media statements

Provenance description and provenance tracking are crucial issues in a wide range of domains, such as Museums, Libraries and Archives (MLA), geoscience, computer science. Provenance has been used to a wide range of domains to identify data trustworthiness, track ownership and/or authorship of works, audit errors, reproduce research data, and so forth. However, the studies conducted for provenance description of metadata in the research area of metadata longevity are quite limited.

3.5. Provenance Related Standards, Models and Vocabularies

From this survey, the author learned that there are already a wide range of models, ontologies and vocabularies that can be used for provenance description. The Provenance Working Group at W3C has published PROV family of documents, including the PROV Data Model (PROV-DM), PROV Ontology (PROV-O), and so forth. The working group aims at the description and interoperable interchange of provenance information in heterogeneous environments such as the Web. PROV-DM is a conceptual data model, which defines a set of concepts and relations to represent provenance (Moreau et al., 2013). PROV-O defines a set of classes and properties as an OWL2 ontology allowing mapping PROV-DM to RDF (Lebo et al., 2013). PROV-DM is derived from Open Provenance Model (OPM).

OPM is a research result of the International Provenance and Annotation Workshop (IPAW). Based on the OPM Core Specification (v1.1), the OPM is designed to meet six requirements, such as exchange of provenance information between systems, representation of provenance for any "thing", and so forth (Moreau et al., 2010). OPM Vocabulary (OPMV), OPM OWL Ontology (OPMO) and OPM for Workflows (OPMW) are defined pertaining to OPM. OPMV as an OWL-DL ontology is designed to assist the interoperability of provenance information on the Semantic Web and to support provenance descriptions for datasets beyond those in the Web of Data (Zhao,

2010). OPMO as an OWL ontology allows full expressivity of OPM concepts and supports inferencing (Moreau et al., 2010). OPMW is an OWL-DL ontology developed to represent abstract workflows and workflow execution traces. OPMW extends and reuses OPM's core ontologies. In the latest release, OPMW also extends PROV to represent scientific processes (Garijo and Gil, 2014).

W7 model is developed to represent the semantics of data provenance in which provenance is conceptualized as a combination of seven interconnected elements including "what (occurring event)", "how (action leading to event)", "who (involved individuals or organizations)", "when (time of event)", "where (location of event)", "which (software or instrument that was used)" and "why (reason for why event happened)" (Liu, 2011). A Vocabulary for Data and Dataset Provenance (Voidp) defines terms to describe provenance relationships of data in linked datasets (Omitola et al., 2011). Provenance Vocabulary (PRV) as an OWL-DL ontology defines classes and properties for describing provenance of Linked Data on the Web. PRV is a domain specific specialization of PROV-O. It is notable that PRV defines terms for both data creation and data access (Hartig and Zhao, 2012). Provenance, Authoring and Versioning Ontology (PAV) is designed for the capture of essential descriptions for tracking the provenance, authoring and versioning of Web resources (Ciccarese et al., 2013). BBC Provenance Ontology is designed to capture data about the provenance of data in an RDF Triple Store (BBC, 2012). Provenir Ontology (PO) defined in OWL-DL defines classes and properties to represent provenance metadata in eScience (Sahoo and Sheth, 2009).

The author analyzed the existing provenance description models and vocabularies (Li and Sugimoto, 2014) and learned that: (1) some models are general and can be tuned to specific domains, for example, PROV data model, Open Provenance Model; (2) some are designed to specific applications, for instance, BBC Provenance Ontology. The existing models do not cover description of structural and semantic features of metadata. In other words, those models lack classes and properties defined for describing changes in metadata application profiles and metadata vocabularies. Therefore, this study analyzes requirements to describe revision history of metadata application profiles and metadata vocabularies through provenance modeling.

W3C PROV standard refers to various aspects of provenance, such as modeling, serialization, exchange, access, validation, semantics and reasoning (Moreau et al., 2015). W3C PROV defines a set of specifications, for instance, a conceptual data model (PROV-DM) and an OWL Ontology (PROV-O) for provenance description and interchange. W3C PROV has been applied to represent provenance description of geospatial objects and global change information (Masó et al., 2015; Tilmes et al., 2013). Missier and Chen (2013) encoded revision history of wiki pages using PROV-

DM. ProvONE conceptual data model developed in DataONE project is for provenance representation of scientific workflow by extending W3C PROV (Cuevas-Vicenttín et al., 2015; Missier et al., 2013). Lagoze et al. (2013) applied PROV-DM to social science data with use scenarios of provenance descriptions. The Oxford University Research Archive (ORA) describes digital objects with contextual information and provenance of scholarly outputs using ORA data model, which was devised by incorporating PROV-DM for activity description (Jones et al., 2015).

Many provenance related models, vocabularies and ontologies have been developed to describe provenance for general or specific application. Previous studies have shown that W3C PROV is commonly applied in several domains because of its strong extendibility. However, those applications of W3C PROV are not specialized for metadata schema. This study adopts and applies W3C PROV to describe metadata provenance especially focusing on provenance description of metadata application profiles and metadata vocabularies.

3.6. Provenance Tracking and Representation of Changes

Provenance, context, and lineage are key components in data curation (Mayernik et al., 2013). Shaon (2006) briefly presented requirements of long-term metadata curation including metadata preservation, metadata quality assurance, metadata versioning, provenance tracking, and other aspects. “Metadata curation record” was proposed to document changes of a digital object throughout its lifecycle and associated metadata. High quality metadata is significant to successful long-term preservation (Shaon, 2005). Sousa et al. (2014) discussed assessment of metadata quality using provenance for long-term accessibility of scientific data. Factor et al. (2009) addressed the importance of provenance (history of creation, ownership, accesses, and changes of digital objects) for long-term use of digital objects and proposed to document provenance as chronologically ordered records describing the events over the lifecycle of content data. Auer et al. (2012) and Papastefanatos (2014) stated issues about long-term accessibility of LOD referring to provenance tracking, change detection and multi-version archiving. Although these previous research papers have reported that provenance tracking is quite significant to long-term use of digital objects, the efforts to metadata provenance for metadata longevity are yet sufficiently undertaken. Hence, this study presents model proposals of metadata provenance description for long-term maintenance of metadata schemas.

Javed et al. (2014) proposed a layered change log model to record the changes of ontology using RDF triple-based representation. Chawuthai et al. (2016) presented a logical model named Linked Taxonomic Knowledge (LTK) and LTK Ontology for preserving and representing changes in taxonomic knowledge for Linked Data. The changes in conception or in the relationship between

taxa are preserved as events along with aspects of time, provenance, causes, and effects. Changeset vocabulary defines a set of terms (for example, Addition, ChangeReason, and Removal) to describe changes between two versions of a resource description by using two sets of triples, i.e., additions and removals (Tunnicliffe and Davis, 2009). Changeset vocabulary represents changes to resource descriptions using RDF reification. An update is represented by a set of statements about statements and whether they are added or removed (Meinhardt, 2015). Changeset vocabulary is used by LCSH to describe the information of "Change Notes" of subject headings. The document-centric approved list of new headings and revisions to existing headings in LCSH are available on the Acquisitions and Bibliographic Access Web page. The changes to the subject headings are provided together with the literal words like "ADD FIELD" or "DELETE FIELD". Although Changeset vocabulary is applicable to describe changes of metadata vocabularies, the use of RDF reification makes the description of changes of metadata vocabularies complex.

Sompel et al. (2010) proposed a versioning mechanism based on the Memento framework and applied the versioning approach to Linked Data. Halpin and Cheney (2014) discussed changes in dynamic RDF datasets over time. They explored the ways to represent provenance records as RDF using named graphs and provide provenance information as a SPARQL query. The changes are recorded using their own change metadata ontology and existing Provenance Vocabulary Core Ontology terms. A tool supporting version management of RDF vocabularies named SemVersion has been developed (Kendall et al., 2008). Meinhardt (2015) presented a model for linked datasets and their evolution and proposed a service approach to preserve the history of linked datasets (Meinhardt, 2015). SemVersion provides structural and semantic versioning for RDF models and RDF-based ontology language like RDFS (Völkel and Groza, 2006). The data model PROV-DM defined by W3C PROV is used to encode the revision history of wiki pages (Missier and Chen, 2013). Getty Thesaurus of Geographic Names adopts W3C PROV to describe revision history of geographic names. W3C PROV has also been used to document the Activity information about the revision of geographic names, e.g., Activity type (Create, Modify) and temporal information associated with the Activity.

Previous studies have proposed approaches to record changes in ontology, taxa and RDF datasets. However, this study is different from the above studies because it mainly focuses on structural changes in metadata application profiles and semantic changes in metadata vocabularies based on the Singapore Framework of DCAP.

3.7. Provenance Usage in the Libraries, Archives and Museums

3.7.1. Provenance Usage in the Library Community

Library of Congress launched Audio-Visual Prototyping Project from 1999 to 2004 for digital preservation of audio and video. The project achieved Digital Production and Provenance Metadata Extension Schema (DIGPROVMD)[11] that is used to document a digital production process. The digital production process is defined as the people, methods, activities, and infrastructure involved in the conservation treatments and the digitization of the archival object. DIGIPROVMD defines five top-level elements including process, task, tool, settings and configuration file. The specific elements are defined in DigiProv Data Dictionary.[12] The achievements of the project are kind of provenance initiative relevant to preservation issues in the library community. Recent practices related to provenance referring to identifying ownership of rare books using provenance, describing research outputs with provenance information, and publishing bibliographic data with provenance information as Linked Data (Kumar et al., 2013).

Provenance information about library collections indicates association of collections and can be used to identify authentication of a rare book, e.g., if the rare book is a spurious work or a facsimile work. The CERL provides records of rare books including provenance information. User can search owners of rare books using the CERL website. CERL has added a Provenance Names section to the CERL Thesaurus.

The Bodleian libraries at University of Oxford devised a data model to represent contextual information of research outputs in the Oxford University Research Archive (ORA), which is a long-term data repository for scholarly research outputs. The model incorporates PROV-DM to describe activity related to research outputs, e.g., creation activity, funding activity, publication activity. Activity-based description of relationships for a journal article using PROV-O is given as an example (Burgess, 2016).

Provenance in the library domain is becoming increasingly important, especially when library data is published as Linked Data. Provenance of library data should be provided on the Web to indicate the source of library data and derivation relationships between different data sources. How to use provenance to reveal the trust and quality of bibliographic data needs further research (Kumar et al., 2013).

[11] Please see http://lcweb2.loc.gov/mets/Schemas/PMD.xsd
[12] Please see http://www.loc.gov/rr/mopic/avprot/DD_PMD.html

3.7.2. Provenance Usage in the Archival Community

In the archival domain, provenance is mainly used to arrange archives, provide contextual information of digital archives, and ensure trustworthiness of digital records.

According to the Society of American Archivists glossary, provenance is the fundamental principle of archives, referring to the individual, family or organization that created or received the items in a collection. The records are arranged through the retention of original order and their placement in their original collections based on their provenance information (Pearce-Moses, 2005). Capturing the provenance information of electronic records and keeping provenance of the archived items are concerns for archivists. Conventional provenance in the arrangement of archival records are creators, for example, individuals, cooperated bodies or families. The scope of provenance for archival records encompasses to creator history, records history and custodial history.

The archival standards mentioned in Section 3.4, such as, ISAD(G), EAD, ISAAR(CPF) and EAC define the description elements for provenance information. The recordkeeping metadata standard ISO 23081-1 (2017) provides guidance to capture audit trails in the records management process. Provenance information provides evidence for authenticity of electronic records over their lifecycle, which is also addressed by InterPARES project.

3.7.3. Provenance Usage in the Museum Community

Provenance description about the history of ownership, custody, and movement of art is critical for understanding the events, people, and locations that are significant to the history of an object. Many museums provide provenance information of paintings on the Web, e.g., Carnegie Museum of Art and Indianapolis museum of Art. Cleveland museum of Art provides provenance research database to enable users to search provenance. The Getty Provenance Index Databases provide search services for its archival inventories, sales catalogs, and provenance of paintings.

On the Web, there is a need to represent the provenance of works as structured data for both computers and humans. Usually provenance is recorded in a text field within collections management system. “Art Tracks: Standardizing digital provenance documentation for cultural objects” project[13] has been conducted to create a digital model for storing and capturing data with provenance in a machine-readable format. The project was launched in early 2013 and established the Carnegie Museum of Art (CMOA) digital provenance standard. The standard is for digitizing and serializing provenance, bridging the gap between the traditional human form of provenance

[13] Please see http://www.museumprovenance.org/reference/standard/

records and the future paradigm of LOD. The standard has three expressions that are compatible and can be converted to each other. The first is a textual model as an extension of the American Alliance of Museums (AAM) recommended provenance text. The second is as an abstract data model in JSON. The third is as a LOD, as a RDF model using the CIDOC-CRM (The CMOA Digital Provenance Standard, 2016). In addition to the standard, the Art Tracks project also achieved the development of software, e.g., museum-provenance library and Elysa. The museum-provenance library support conversion between textual provenance record using CMOA provenance model and structured data. Elysa is a Web-based user interface to assist in reviewing and modifying provenance records.

In addition, the widely used CIDOC-CRM has been extended to model provenance information of digital objects (Theodoridou et al., 2010). CRMdig ontology has been developed as an extension of CIDOC-CRM to support provenance metadata. CRMdig declares a set of classes and properties to encode metadata about the steps and methods of production ("provenance") of digitization products and synthetic digital representations.

It is important for memory institutions to record and provide provenance information of their holdings. W3C Provenance Incubator Group (2010) listed provenance-related use cases, which include provenance in cultural heritage. Europeana provides access to resources held at cultural heritage institutions throughout Europe. Europeana is a use case of metadata provenance, in which metadata provenance is represented via Europeana Data Model using OAI-ORE model (Eckert, 2012).

As introduced above, provenance description is necessary for both physical and digital collections that are managed and preserved in MLA. MLA communities have paid attention to provenance description, especially the change history and activity related to objects. There are models and standards referring to provenance description. However, these provenance description elements are designed for specific domain requirements and not generalized for metadata provenance. That is, they cannot be directly applied to describe provenance of metadata application profiles and metadata vocabularies. Therefore, the aim of this research to propose general models for provenance description of metadata is novel.

Furthermore, provenance provision and query services are limited, especially the cases working together with Semantic Web technologies. With the development and growth of Linked Data services, provenance description of digital objects and metadata objects in machine-understandable form is required on the Semantic Web.

3.8. Provenance in the Web Environment

Moreau (2010) gave a comprehensive introduction and review about provenance and provenance research. Moreau summarized definitions of provenance from dictionaries to its alternative definitions, such as “provenance as process”, “provenance as a directed acyclic graph”, “why provenance”, “where provenance”, “how provenance”, “provenance as annotations”. Moreau summarized provenance related literatures referring to six clusters, i.e., databases, workflows, eScience, Provenance Challenge, Open Provenance Model, Semantic Web and Accountability. Moreau’s work lays a foundation for the provenance and contributes a lot to the development of OPM and W3C PROV. His studies about provenance are mainly from perspectives of computer science. This study focuses on provenance in the metadata community.

Publishing provenance as LOD and provenance use in data quality has been discussed. Hartig and Zhao (2010) presented an approach to publish the provenance of structured data on the Web as LOD via using their own developed provenance vocabulary and existing widely used LOD publishing tools. They published provenance metadata to make them accessible and applied provenance for timeliness comparison to identify outdated information in specific gene data. Omitola et al. (2011) presented voidp including classes and properties (a provenance extension of voiD vocabulary) for data publishers to provide provenance. In addition, they carried out the experiment using United Kingdom’s public data as a use case scenario. Sharma et al. (2014) gave an overall review of the approaches for converting legacy data to LOD together with provenance tracking (referring to provenance type, provenance model, and provenance storage) over the LOD generation. Anam et al. (2015) distinguished Linked Data provenance into instance level and schema level. They mainly presented how provenance information about schema level mapping can be represented, stored and queried. Flouris et al. (2012) stated provenance is a critical factor for data quality assessment. They proposed quality metrics considering provenance and extended existing approaches for LOD datasets repairing.

The Web has become a global information space and the Semantic Web facilitates the forming of a global web of Linked Data. Semantic Web technologies (e.g., RDF and SARQL) can be used to represent, query and reason provenance. Metadata exchanged on the Web plays important roles in sharing and interchange of data, provenance tracking of RDF-based metadata is required on the Web. Given to provenance roles in data quality and data trust, provenance of metadata can provide useful evidence for data quality judgments. Provenance description is required in both conventional and Web environment. However, existing technologies and standards are not specialized for metadata schema and metadata vocabulary. Specially, models for formal provenance description of metadata are not sufficiently explored. On the Semantic Web, there is a need to develop models

for formal metadata provenance description interpretable by both computers and humans. It is because that formal provenance description of metadata in machine-readable and interoperable form supports automated and effective metadata maintenance. In this study, the author has developed models for formal provenance description of metadata application profiles and metadata vocabularies, respectively.

4. Provenance Description using PROV with PREMIS

Metadata should be preserved as well as primary resources to keep the primary resources alive. Metadata preservation is important as well as preservation of primary digital resources. Metadata preservation is an important research topic for keeping metadata about preserved resources consistently usable over time. Provenance information is necessary for long-term use and preservation of digital resources. Provenance about metadata objects should be recorded for metadata longevity over time.

This chapter focuses on provenance as an important issue in both digital preservation and metadata preservation, which discusses provenance description based on two major metadata standards – PROV and PREMIS. Implementation of existing provenance models with metadata standards (e.g., PREMIS dictionary; controlled vocabularies of Library of Congress) is an applicable approach for provenance description of metadata. This chapter used this approach and briefly discussed provenance description of metadata schemas through combining the core of PROV data model with PREMIS data model. This chapter first introduces well-known standards – OAIS, PREMIS, PROV, and then discusses provenance description based on the PROV Ontology (PROV-O) and PREMIS OWL Ontology with examples. Based on analysis and mapping among the basic classes of the PROV-O and PREMIS OWL Ontology, the author proposes an approach of describing provenance for metadata preservation by integrating PROV-O with PREMIS OWL Ontology.

4.1. Digital Provenance in OAIS and PREMIS

Provenance description is necessary for long-term preservation of digital resources. Open Archival Information System (OAIS) and Preservation Metadata: Implementation Strategies (PREMIS), which are well-known standards designed for digital preservation, define descriptive elements for digital preservation.

The Open Archival Information System (OAIS) defines three metadata components, which have to be maintained with Digital Object – Representation Information of Digital Object, Preservation Description Information (PDI) in an Information Package, and the Content Information given to every Information Package. Provenance of a digital object, which is one of the five categories of PDI, is a crucial record of the history of the object over its lifecycle. Those metadata may be stored in a database with the preserved digital objects as an Archival Information Package (AIP). This means that metadata schemas and vocabularies used in those metadata have

to be maintained over time as well as those AIPs to keep the information object interpretable, i.e., renderable, playable, operable, and functional in various ways.

The OAIS reference model is a widely used model for archiving and preserving digital resources. Digital objects are preserved as a sequence of bits. It is of importance to ensure that the bits remain intact and correct over time. However, bit preservation alone is not sufficient for the long-term preservation of digital objects. Digital objects should be kept interpretable across the changes in many aspects over time. Although the Information Package model in OAIS does not address metadata longevity very well, it provides important insights for the longevity of digital objects.

Provenance information in OAIS is defined as the history of the Content Information, which describes the origin of and changes on an archived resource, and agents who hold custody since its origination (CCSDS, 2012). The provenance description is a part of PDI, and documents evolutionary processing history associated with the Content Information over its complete life cycle.

PREMIS is a widely used international metadata standard for the preservation of digital objects. The PREMIS Data Model defines five Entities for digital preservation, which are Intellectual Entity, (Digital) Object, Event, Agent, and Right (PREMIS Data Dictionary Version 3.0, 2015). Documentation of actions on a digital object is critical for the maintenance of the object. The documentation, i.e., metadata about the actions, is aggregated as an Event. Thus, Event is crucial component for provenance description associated with Object. PREMIS Data Dictionary defines a set of descriptive elements of the five Entities. Those elements are called semantic units. Some of the semantic units associated with an Event record changes to a preserved digital object (PREMIS Editorial Committee, 2012). PREMIS OWL Ontology defines classes and properties to describe preservation metadata in RDF. Provenance may be about any resource, such as documents, rare books, Web pages, datasets, transaction execution records, etc. The use of appropriate vocabulary (-ies) for provenance description should be in accordance with the type of resources and archiving purposes. Provenance description in OAIS and PREMIS is primarily for digital preservation.

Digital Provenance is chronology or chronological information related to management of a digital object. Digital provenance typically describes agents responsible for the custody and stewardship of digital objects, key events that occur over the course of the digital object's life cycle, and other information associated with the digital object's creation, management, and preservation (PREMIS Editorial Committee, 2012).

The following sections show some cases of provenance description about the format migration, referring to the *generationActivity/creationEvent* occurred to Digital Object A,

responsible Agent, related date time, and the derivation of Digital Object A in Format X to Digital Object B in Format Y via *migrationActivity* which caused the format change, and so forth.

4.1.1. Description of Activity and Event

Figure 4.1 shows a *generationActivity* leading to the generation of Object A by using PROV. The generationActivity (started at dateTime1, ended at dateTime2) resource is directed to Object A, which is linked to a generation Date-Time literal. PREMIS uses preservation-specific value vocabularies defined by Library of Congress. The controlled values are expressed in SKOS vocabularies for *EventType*, *AgentType*, *RelationshipType*, and so forth. Likewise, Figure 4.2 shows a *creationEvent* associated with Object A and the creationEvent happening during a period from dateTime1 to dateTime2. Meanwhile, the Figure also presents the creationEvent is linked to an *EventOutcomeInformation* resource, an *EventType* resource, and *EventDateTime* literal.

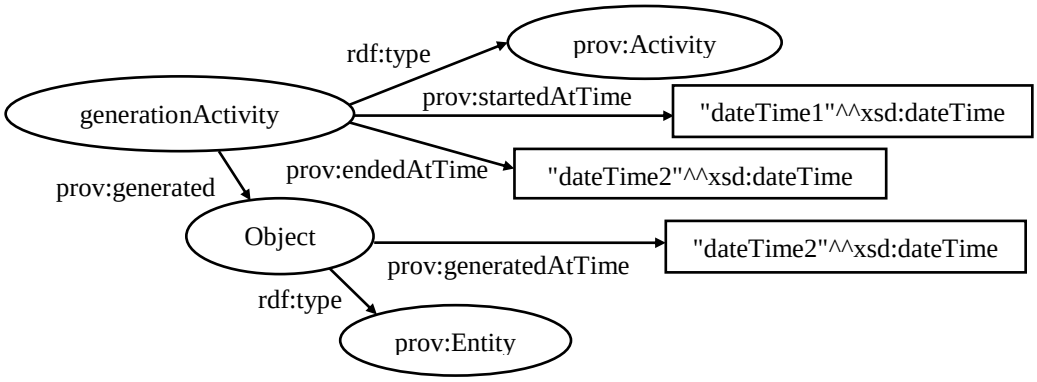

Figure 4.1: Provenance graph of generationActivity happened on Digital Object A using PROV.

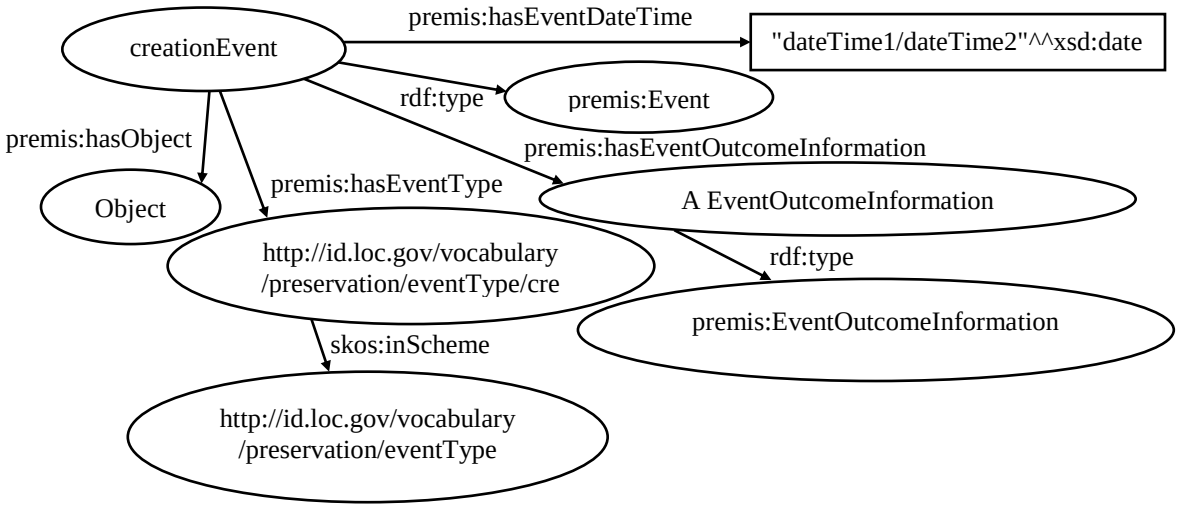

Figure 4.2: Provenance graph of creationEvent occurred to Digital Object A using PREMIS.

4.1.2. Description of Responsible Agent

As shown in Figure 4.3, Object A is connected with a Person by property *wasAttributedTo* defined in PROV. The *generationAcitity* is linked to that Person via property *wasAssociatedWith*, from which we know the Person holds a responsibility for the generation of Object A. In PREMIS, Agent influences Object through Event. That is, Agent is not directly connected to Object as shown in Figure 4.4. However, PROV allows Agent, Entity and Activity to be related with each other directly. In PREMIS, Agent influences Object through Event. That is, Agent is not directly connected to Object as shown in Figure 4.4. However, PROV allows Agent, Entity and Activity to be related with each other directly.

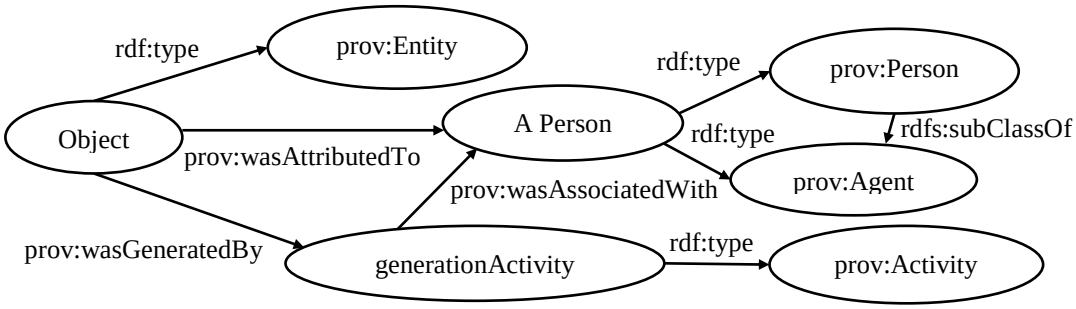

Figure 4.3: Provenance graph of Agent responsible for the generation of Digital Object A Using PROV.

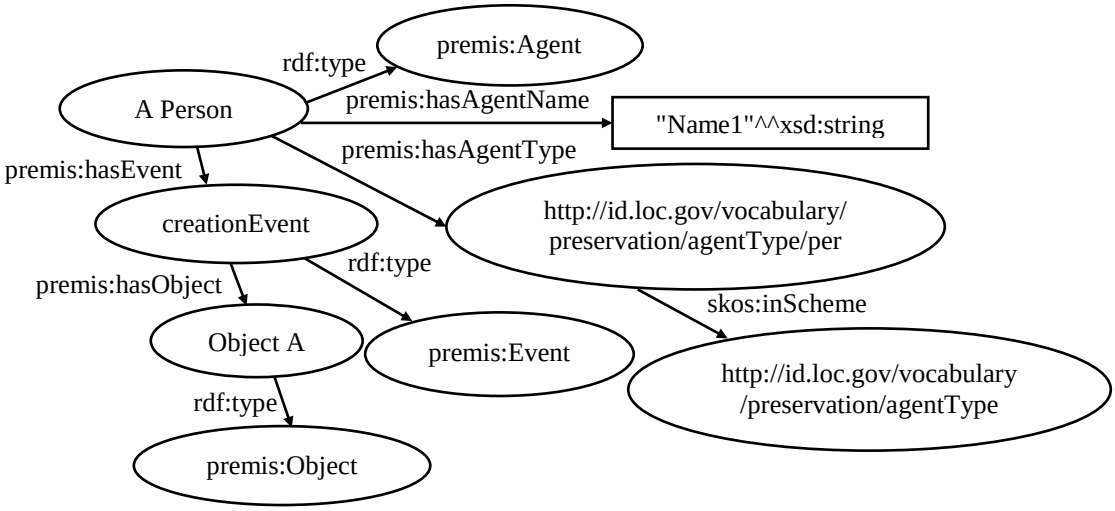

Figure 4.4: Provenance graph of Agent responsible for Event using PREMIS.

4.1.3. Description of Relationships between Entities and Objects

PROV defines the relationships between Entities using properties *wasDerivedFrom*, *alternateOf*, *specializationOf*, *wasQuotedFrom*, *wasRevisionOf*, *hadPrimarySource*, and *hadMember*. Figure 4.5 shows that Object A is the primary source of Object B using PROV. PREMIS holds two types of relationship between Objects, including structural relationship and derivation relationship defined in SKOS vocabulary by Library of Congress. Using PREMIS, Figure 4.6 shows the derivation relationship between Object A and Object B due to the *migrationActivity*.

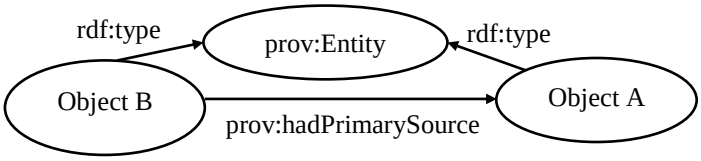

Figure 4.5: Derivation Relationship between Digital Object A and Digital Object B using PROV.

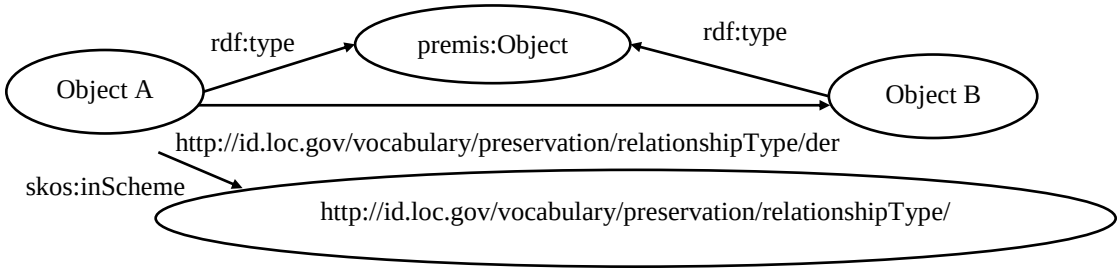

Figure 4.6: Derivation relationship between Digital Object A and Digital Object B using PREMIS.

Furthermore, PROV also defines relationships between Activities and relationships between Agents, whereas PREMIS does not include those relationships. Figure 4.7 shows the relationship expressed by property *wasInformedBy* between the *migrationActivity* and *generationActivity*, which means the *migrationActivity* used Object A created by the *generationActivity*.

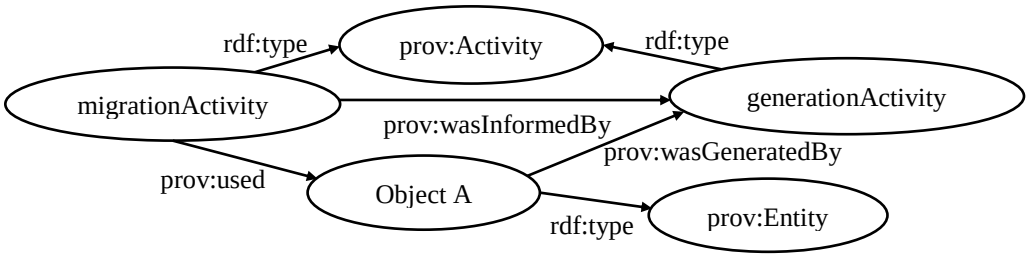

Figure 4.7: Relationship between Activities in PROV.

4.2. Metadata Provenance based on PROV with PREMIS

PROV is designed generally and comprehensively for provenance description, referring to representation, interchange, query, access, and validation of provenance. PREMIS is widely used for digital preservation where provenance description is an important component. PROV and PREMIS are used as a basis for general provenance description and provenance description for preservation.

PROV-O and PREMIS OWL Ontology are used to describe provenance information created in a lifecycle of digital objects and their metadata. For convenience, the author writes PROV and PREMIS instead of PROV-O and PREMIS OWL Ontology in the following sections unless there is a need to explicitly state ontology.

4.2.1. Mapping of the Basic Classes between PROV-O and PREMIS OWL Ontology

PROV has the three base classes, i.e., *prov:Entity*, *prov:Agent* and *prov:Activity*. PREMIS defines classes, including *premis:IntellectualEntity*, *premis:Object*, *premis:Agent*, *premis:Event*, and so forth. Based on the interpretation in PROV (Lebo et al., 2013) and PREMIS (PREMIS Editorial Committee, 2012), the paragraphs below discuss mappings between them.

premis:IntellectualEntity is a set of content items as a single intellectual unit, e.g., book, map, photograph, or database. *premis:Object* is a discrete unit of information in digital form. *prov:Entity* can be in physical or digital or conceptual or imaginary thing. Therefore, *prov:Entity* has a broader meaning than *premis:IntellectualEntity* and *premis:Object*. Hence, the author maps *premis:IntellectualEntity* and *premis:Object* as subclass of *prov:Entity*.

premis:Event indicates a description about an action (or activity) impacting an Object. *prov:Activity* means actions or processes performed by Agent(s) or acted on Entity (-ies). *premis:Event* is oriented to preservation actions, and only important Events are recorded. On the other hand, *prov:Activity* does not have limitation of action domain or types. That is, the meaning of *premis:Event* is narrower than *prov:Activity*. Therefore, the author maps *premis:Event* as subclass of *prov:Activity*.

premis:Agent can be a person, or an organization, or a software program/system associated with Events in the life of an Object. *prov:Agent* bears responsibility for occurred Activity, or the existence of Entity. However, their Agent types are almost the same. *premis:Agent* can be seen to be equal to *prov:Agent*. And the relationship between them can be described using *owl:equivalentClass*.

4.2.2. A Merged Model by Integrating PROV-O with PREMIS OWL Ontology

Both PROV and PREMIS have properties to describe provenance, and they are defined based on RDF and OWL. PROV is designed for generalized provenance description and interchange among different systems, whereas PREMIS is primarily for preservation metadata description used for digital preservation. The PREMIS terms used to describe preservation could enrich expressive power of PROV. By introducing the controlled vocabulary for *EventType* suggested in PREMIS, interoperability of Activity descriptions in PROV could be enhanced.

Based on the above mapping, the author proposes a provenance description model for preservation of digital resources and metadata, by integrating the PROV with PREMIS. The merged model shown in Figure 4.8 introduces the *premis:Object* and *premis:IntellectualEntity* as the subclass of *prov:Entity*, Collection, Bundle, and Plan are also subclasses of Entity. Meanwhile, *premis:Event* is mapped to the subclass of *prov:Activity*, *premis:Agent* is equivalent to *prov:Agent*. In the Figure, the classes in PROV are written in italic, and the classes in PREMIS are shown with underline. Moreover, as shown in Figure 4.8, the relationships between classes, the generation or invalidation time of Entity, and the start or end time of Activity/Event can also be described using properties (written with namespace prefix, i.e., prov) from PROV.

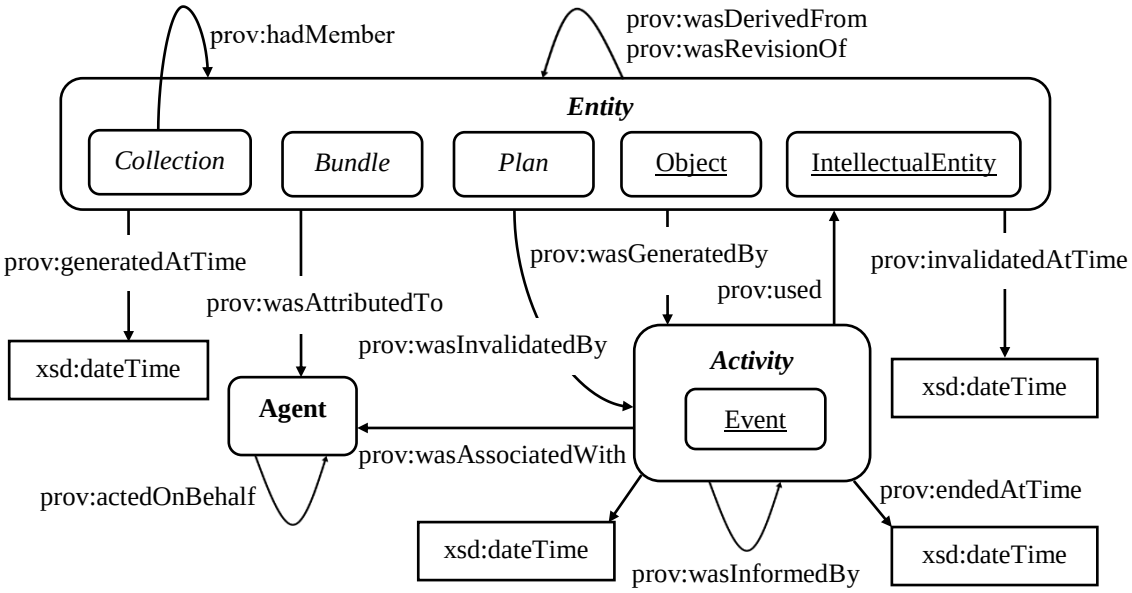

Figure 4.8: The merged model for provenance description oriented to digital preservation.

4.2.3. Metadata Provenance Description Example

Eckert presented the concept of Provenance Context. A Provenance Context can be seen as a Named Graph about identified resource (Eckert, 2013). Named Graph may be used for tracking provenance of RDF data, replication of RDF graphs, and versioning (Dodds and Davis, 2012). PROV allows grouping of provenance description and defines Bundle as a named set of descriptions (Lebo et al., 2013).

Through the definition of Bundle, the provenance of Bundle can be described. In the example shown in Figure 4.9, Digital Object A in Format X is migrated to Digital Object B in Format Y. Here, the author defines two Bundles, i.e., *Bundle 1* and *Bundle 2*. *Bundle 1* and *Bundle 2* respectively describes the format feature of Digital Object A and Digital Object B as shown in Figure 4.9, which shows the format change caused by *migrationActivity*. As Bundle is an Entity in PROV, we can also express the derivation between *Bundle 1* and *Bundle 2*. In PROV, by using property *qualifiedDerivation*, we can qualify how *Bundle 2* was derived from *Bundle 1*. In Figure 4.9, *Bundle 2* is linked to a blank node through property *qualifiedDerivation*. And from the blank node, the *migrationActivity* caused the format change is expressed.

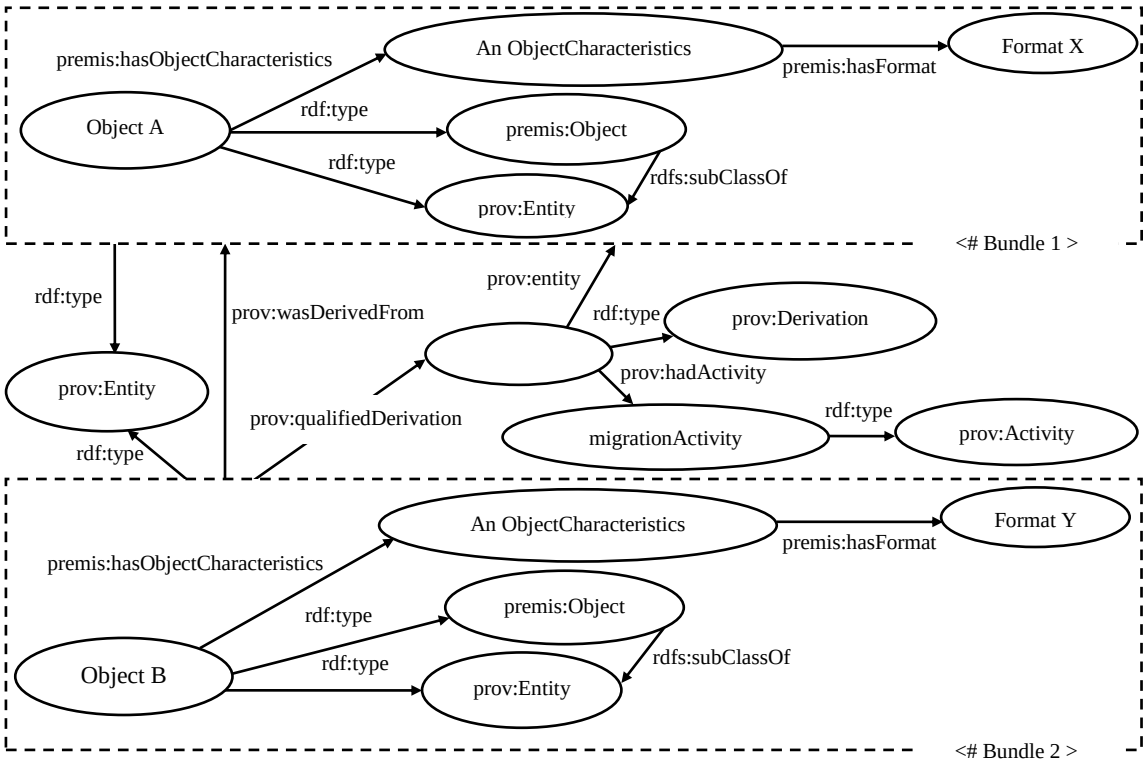

Figure 4.9: Provenance graph of the format change from Digital Object A to B using Bundle.

4.3. Summary

In this chapter, the author addressed provenance description of metadata using PREMIS standard for preservation and W3C PROV standard for provenance. The merged model was defined as a primary model by mapping the core classes of PREMIS ontology and W3C PROV-O. The merger of PREMIS and PROV combines the perspectives from both preservation and provenance. W3C PROV "provides an avenue for handling provenance of digital objects including metadata and metadata schema" (Haynes, 2018). The incorporation of preservation perspectives to PROV for provenance description of metadata brings in characteristics of metadata as digital objects, which is helpful to examine key events and change history of metadata over time.

Both PREMIS ontology and PROV-O have a large set of classes and properties. It is a huge work to propose a comprehensive provenance model through merger of the whole sets of PREMIS ontology and PROV-O. Therefore, in this chapter, the author provided a core model for metadata provenance based on PREMIS and PROV. This approach would assist in the description of provenance of metadata, such as who created the metadata, what rules were used to create it, and when was it created or amended (Haynes, 2018). The primary model proposed in this study can be implemented for archival services such as archival systems that need provenance description for the long-term use of digital objects.

5. Provenance for Long-term Maintenance of Metadata Schema

In conventional systems, since metadata in conventional services has been mostly organized as a database, maintenance of the metadata is likely to be recognized as maintenance of the database. In such environment, the schemas of the metadata are documented as a part of the database schema. Those schema documents are maintained primarily for human-readers. The author considers that this is the main reason of the lack of research on long-term maintenance of metadata schemas. However, in the state-of-the-art Web environment today so called LOD environment, there is a need of metadata schema maintenance technologies drastically different from that used in the conventional database-centric environment. This is because both metadata and their schemas can be encoded in XML and transferred from a site to another as a first-class object in the LOD environment. Sugimoto et al. (2016) presented differences between conventional and LOD environment for metadata schema maintenance and discussed facets in long-term maintenance of metadata schemas in the LOD environment. Long-term maintenance of metadata schemas in the LOD environment need to use the technologies that fit to LOD but are not well developed yet.

The author has learned the importance of provenance description of metadata schemas from Preservation Description Information (PDI) of OAIS. Among the five categories in PDI, which are Reference, Provenance, Context, Fixity, and Access Rights, the Provenance category is directly related to events which may cause changes in the preserved objects. It is crucial for long-term maintenance of metadata to keep track of changes in their metadata schema as a digital object which should be readable by machines as well as humans. Provenance description of metadata schemas in Resource Description Framework (RDF) is crucial for the longevity of metadata. In this study, the author aims at proposing a model to formally describe provenance of metadata application profiles for automated tracking of their change history and consistent maintenance of metadata over time.

The author analyzed the existing provenance description models and vocabularies (Li and Sugimoto, 2014) and learned that: (1) some models are general and can be tuned to specific domains, for example, PROV data model, Open Provenance Model; (2) some are designed to specific applications, for instance, BBC Provenance Ontology. The existing models do not cover description of structural features of metadata. In other words, those models lack classes and properties defined for describing changes in metadata application profiles. Therefore, the author has analyzed requirements to describe revision history of metadata application profiles and defines a provenance description model for metadata application profiles.

In the long term, changes in metadata schemas may cause inconsistencies and incorrect interpretation of metadata. Hence, provenance that describes revision history of metadata schemas should be appropriately recorded. Provenance description in a natural language is not efficient to track changes among versions of a metadata schema. Provenance description should be formally recorded for machine-readability and traceability to audit inconsistent recording of structural changes of a metadata schema. Structural features of a metadata schema which define data structure, mandatory levels and iteration constraints of description should be consistently maintained over time. The author proposes a formal provenance description model with functions to keep track of structural changes of metadata schemas over time. The proposed model is applied to the Metadata Application Profile of Digital Public Library of America (DPLA MAP)[14] to show the advantage of the model against conventional semi-formal description of change logs of structural features of DPLA MAP.

The Semantic Web and LOD activities encourage us to represent links which connect data instances on the Web in a machine-processable format. The machine-processable provenance can be provided using an ontology recommended to describe provenance information, such as W3C's provenance ontology (Lóscio et al., 2017). The Semantic Web technologies facilitate acquisition and representation of provenance descriptions as well as reasoning based on the formal descriptions in RDF (Moreau, 2010). Thus, the model discussed here is purposed to formal provenance description of metadata application profiles using RDF.

The proposed model named DSP-PROV is developed based on the W3C PROV standard and Singapore Framework for Dublin Core Application Profile (DCAP). Singapore Framework for DCAP is used as a generalized model of a metadata schema for an application and its related components, e.g., metadata vocabularies (Heery and Patel, 2000). This study adopts the W3C PROV for provenance description and defines a set of PROV Activities and Entities to describe structural changes of metadata schemas. The DSP-PROV model defines three functions (i.e., addition, deletion and revision) as PROV Activities to formally describe provenance of structural components of metadata schemas based on DCAP.

5.1. Introduction to Description Set Profile

A Description Set Profile (DSP) formally represents the machine-processable part of a Dublin Core Application Profile (Nilsson et al., 2009). A DSP formulates and describes structural

[14] Please see https://dp.la/info/developers/map/

constraints on a description set (Nilsson, 2008). A DSP can be used to examine if metadata records are valid instances of a MAP (Nilsson et al., 2008).

This study defines a DSP and its components as follows. (1) a DSP consists of zero or more Description Template (DTs), (2) a DT consists of zero or more Statement Template (STs), and (3) a ST consists of zero or more Structural Constraints (SCs). This definition allows for a DSP without any DTs, though such DSP would not exist in a practical metadata schema except while a metadata schema is under development.

Figure 5.1 gives a DSP example of a metadata schema to describe a journal paper. All the constraints to describe "Paper" and "Journal" in the rectangle with solid line constitute the DSP. The author illustrates an example of DT and ST using constraints on "Paper" and "Paper Title", respectively. The constraints to describe "Paper" in the rectangle with dotted line constitute a DT of the DSP. The constraints on the property *"dc:title"* used to describe "Paper Title" in the rectangle with broken line constitute a ST of the DT. SCs defined in the ST define that "Paper Title" of a journal paper must be described in literal using the term *"dc:title"* from Dublin Core Metadata Element Set. Another DT of the DSP and one of its ST can be similarly identified as shown in Figure 5.1.

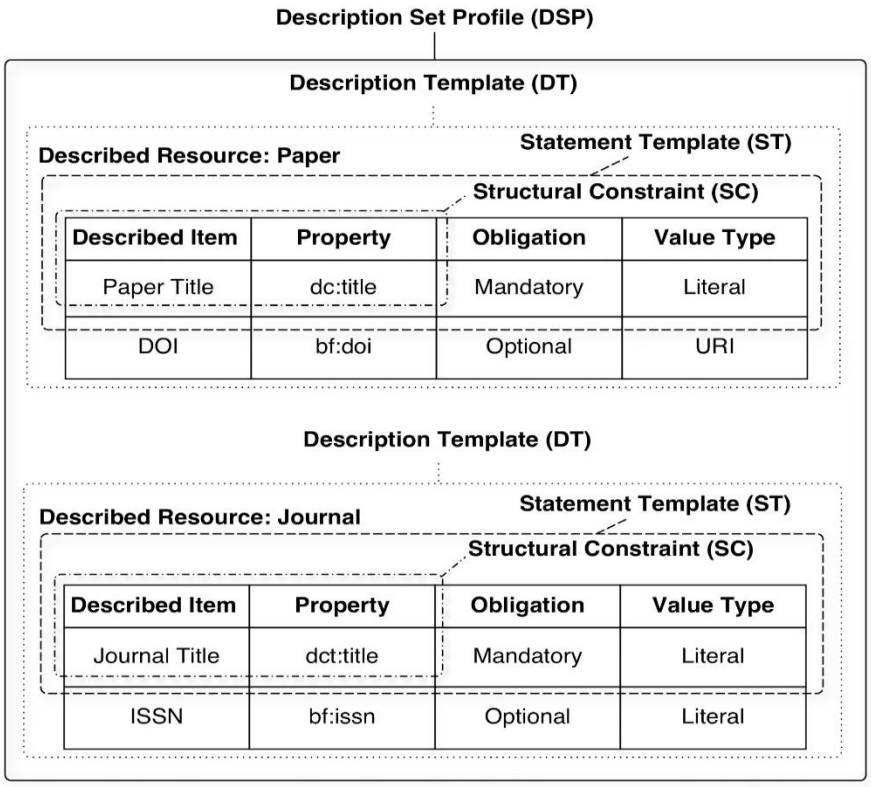

Figure 5.1: Example of a Description Set Profile.

5.2. DSP-PROV Model for Formal Provenance Description of Metadata Application Profile

5.2.1. Classifying Entities to Describe Provenance of Description Set Profile

As stated before, the author applies W3C PROV to describe metadata provenance. According to W3C PROV, Entities and Activities are two important components to describe general provenance. It is necessary to clarify subtypes of Entities and Activities for provenance of Description Set Profile when applying W3C PROV to track structural changes of metadata schemas.

It is straightforward to map Description Set Profile itself and its components as subtypes of Entity because of the broad meaning of PROV Entity. That is, the instances of Description Set Profile, Description Template, Statement Template and Structural Constraint are seen as an instance of PROV Entity. However, it is not straightforward to define Activities influencing structural changes of metadata schema. Structural changes of metadata schemas are caused by Activities acted upon structural schema instances. Therefore, it is required to analyze and classify Activities to describe provenance of Description Set Profile.

5.2.2. Classifying Activities to Describe Provenance of Description Set Profile

Table 5.1 shows a few change documentation in the case of DPLA MAP (DPLA, 2014; DPLA, 2015). These changes are recorded in a semi-controlled style in English.

Table 5.1: Examples of change documentation in DPLA MAP.

Change logs of Digital Public Library of America Metadata Application Profile (DPLA MAP)
Added dpla:intermediateProvider to ore:Aggregation.
Changed obligations for “Collection Title” and “Collection Description” in dc:Collection class.
Deprecation of State Located in property within dpla:sourceResource.
……

The underlined words in Table 5.1 indicate the general change types in DPLA MAP. From the existing documentation, three primitive change patterns in MAPs are extracted and categorized into three actions, deletion, addition and revision. Thus, Deletion, Addition and Revision Activity are defined as primitive Activities to describe provenance of Description Set Profile. Structural

changes of metadata schema are recorded by these three primitive Activities acted upon structural schema instances.

Table 5.2 summarizes the classified Activities to describe structural changes of metadata schema. The naming convention of the Activities in this study is "Activity Type + On + Abbreviation of structural schema instance". For instance, Revision Activity that acted upon a DT and led it to a new DT is named as an activity instance of *RevisionOnDT*.

Table 5.2: Activities to describe structural changes of metadata schema.

Activity	Definition	Description Set Profile (DSP)	Description Template (DT)	Statement Template (ST)	Structural Constraint (SC)
Deletion	Deletion of a DT, ST or SC	–	DeletionOnDT	DeletionOnST	DeletionOnSC
Addition	Addition of a DT, ST or SC	–	AdditionOnDT	AdditionOnST	AdditionOnSC
Revision	Revision of a DSP, DT, ST or SC	RevisionOnDSP	RevisionOnDT	RevisionOnST	RevisionOnSC

5.2.3. Identifying the Relationships among the Classified Activities

The chronological order between the classified Activities is not considered here. Figure 5.2 shows the relationships among classified Activities, which are defined based on the inclusion relationships among structural schema instances.

The Revision Activity acted upon containing Entity (e.g., a DSP) has sub-activities – Deletion, Addition and Revision acted upon its contained Entity (e.g., a DT of the DSP). Changes on a DT caused by *DeletionOnDT*, *AdditionOnDT* and *RevisionOnDT* will result changes in DSP, which in turn specified by *RevisionOnDSP*. Therefore, *RevisionOnDSP* has sub-activities, i.e., *DeletionOnDT*, *AdditionOnDT* and *RevisionOnDT*. Similarly, the following two conclusions are achieved: *RevisionOnDT* has sub-activities, i.e., *DeletionOnST*, *AdditionOnST* and *RevisionOnST*; *RevisionOnST* has sub-activities, i.e., *DeletionOnSC*, *AdditionOnSC* and *RevisionOnSC*. The property *"dcterms:hasPart"* recommended by PROV-FAQ is used here to model sub-activities.

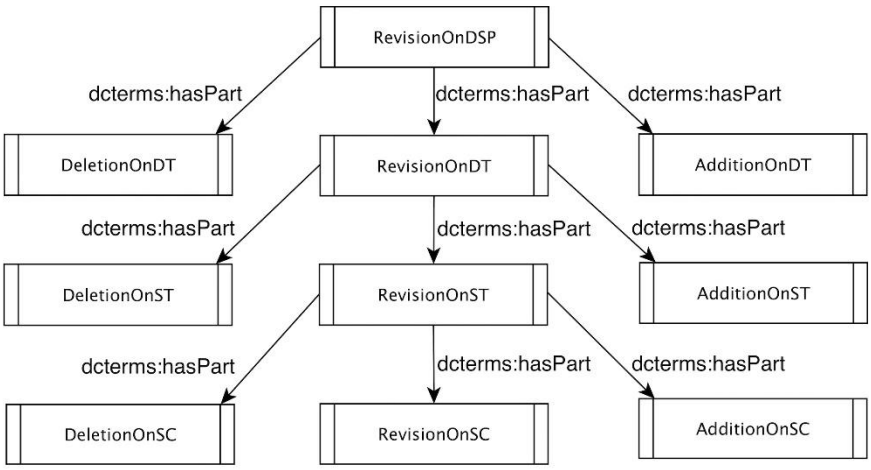

Figure 5.2: Relations among the classified Activities.

Note: DSP, Description Set Profile; DT, Description Template; ST, Statement Template; SC, Structural Constraint.

5.2.4. Overview of DSP-PROV Model

This section shows DSP-PROV model with functions to describe deletion, addition and revision of structural features of a metadata schema. Figure 5.3 depicts the DSP-PROV model using UML Class diagram. (1) Generalization is represented with a hollow triangle on super-classes (i.e., Entity and Activity). (2) Aggregation is represented with a diamond on containing classes (for example, DSP, RevsionOnDSP). (3) Association represented by an arrow describes the relation between an Entity and an Activity.

The DSP-PROV model uses the properties from PROV-O when applicable. PROV Invalidation and PROV Generation respectively represent the deletion and addition of structural schema instances. PROV Derivation, PROV Invalidation, PROV Generation and PROV Usage together describe the revision of structural schema instances. If applicable, DSP-PROV can be also used to describe relations between Activities in the case when an Activity used the Entity generated by another Activity.

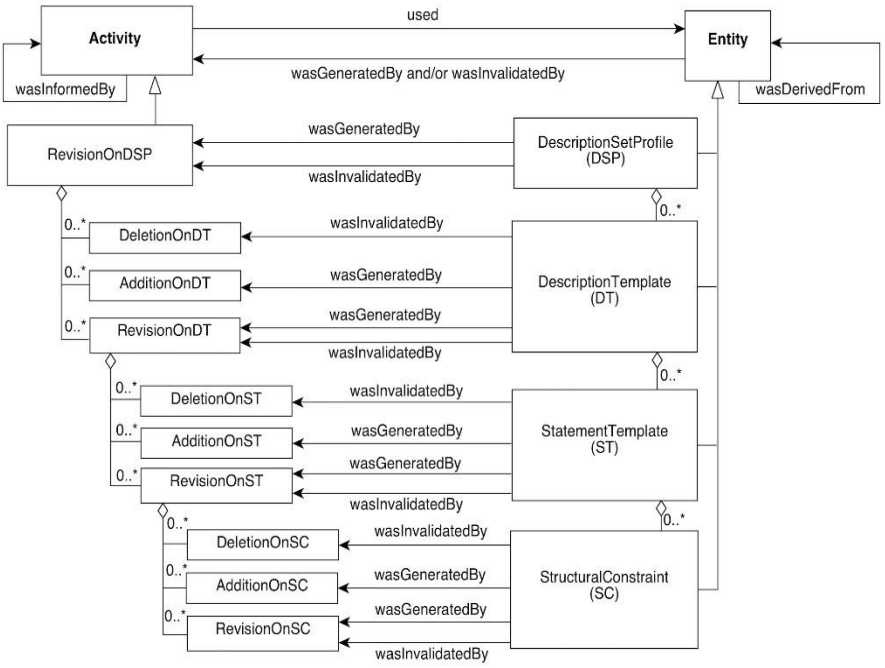

Figure 5.3: DSP-PROV model using UML class diagram.

5.3. Application of DSP-PROV Model to Metadata Application Profile of Digital Public Library of America (DPLA MAP) – A Case Study

5.3.1. Introduction and Selection of DPLA MAP

In this study, the author first collected several documents of metadata application profiles from the projects, such as DPLA, DataCite,[15] CARARE[16] and Dryad.[17] The author first used the following condition to collect the documents, that is, at least two consecutive versions are publicly available on the Web. The author then examined the documents and found that (1) there is no common scheme among these documents, (2) CARARE metadata schema and Dryad application profiles do not provide change logs, (3) provenance descriptions of DPLA metadata application profile (DPLA MAP) and DataCite metadata schema are given as their change logs in pre-defined formats and written in English, which are primarily intended for human readers but not for processing by machines. Next, the author compared DPLA MAP and DataCite metadata schema. DPLA MAP define both classes and properties with namespaces, which can be used to create Description Set Profile of DPLA MAP according to the DCAP. However, DacteCite metadata

[15] Please see https://www.schema.datacite.org/
[16] Please see http://pro.carare.eu/doku.php?id=support:metadata-schema
[17] Please see http://wiki.datadryad.org/Metadata_Profile

schema defines properties without namespaces and declaration of classes. Therefore, the author finally selected DPLA MAP as a case study to apply DSP-PROV model.

DPLA was launched in April 2013 to create a portal for digital collections of America's Libraries, Archives and Museums. We collected three versions of DPLA MAP (V3, V3.1 and V4) that are accessible on the Web (DPLA, 2013; DPLA, 2014; DPLA, 2015). DPLA MAP provides the domain model, usage guide, a set of classes and properties, and change logs between two neighboring versions of DPLA MAP. Table 5.3 shows the definitions of "Collection" class in DPLA MAP V4. In Table 5.3, "Partner-supplied" means the data are supplied by partner of DPLA. "0-1" means the minimum and maximum occurrence of the property.

Table 5.3: Definitions of Class "dcmitype:Collection" in DPLA MAP V4.

Label	Source	Property	Usage	Obligation
CollectionTitle	Partner-supplied	dcterms:title, .sourceResource.collection .title	Name of the collection or aggregation. Literal	0-1
CollectionDescription	Partner-supplied	dcterms:description, .sourceResource.collection .description	Free-text account of aggregation, for example an abstract or content scope note. Literal	0-1

5.3.2. Creation of Description Set Profile of DPLA MAP

Figure 5.4 illustrates the creation process of DSP of DPLA MAP in RDF from DPLA MAP in English. In the first step, the author converted each of the PDF files of the three versions to Excel files using the Nitro[18] online free service. In the second step, the data in the Excel files was manually checked for the next step, e.g., exclusion of non-DSP information, addition of minimum occurrence/maximum occurrence. In the third step, the author imported every version of the corrected Excel data into OpenRefine,[19] and mapped the tabular data to the pre-defined RDF structure, and exported the generated RDF data in Turtle serialization syntax. In the fourth step, Rapper (Raptor RDF Syntax Library)[20] was used to parse the created DSP RDF data for checking their syntactic correctness.

[18] Please see https://www.pdftoexcelonline.com/en/
[19] Please see http://openrefine.org/
[20] Please see http://librdf.org/raptor/rapper.html

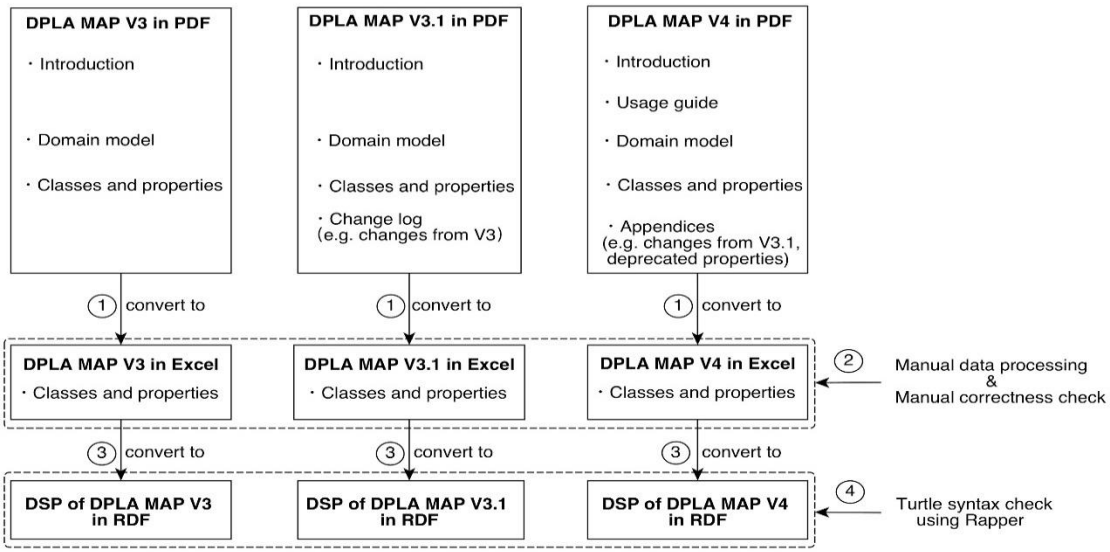

Figure 5.4: Creation of Description Set Profile of DPLA MAP in RDF.

Table 5.4 summarizes the classes and properties used for DSP creation. The classes *"dsp:DescriptionTemplate"* and *"dsp:StatementTemplate"* defined in the vocabulary with namespace *"http://purl.org/metainfo/terms/dsp#"* respectively describe instances of Description Template (DT) and Statement Template (ST). The property *"dsp:subClassOf"* means that a DT has all the common constraints of its contained STs. The properties *"owl:minQualifiedCardinality"* and *"owl:maxQualifiedCardinality"* from OWL 2 Web Ontology Language for Semantic Web are used to describe qualified cardinality restrictions.

Table 5.4: Classes and properties used for Description Set Profile creation.

Class/Property	Definition
dsp:DescriptionTemplate	Is defined as subclass of owl:Class. Puts constraints on instances of a certain described resource class.
dsp:StatementTemplate	Is defined as subclass of owl:Restriction. Puts constraints on every single described item.
dsp:resourceClass	Is defined to represent the belonging resource class of a description template.
rdfs:subClassOf	Is to connect a description template and its contained statement templates.
owl:onProperty	Its value is the used property representing the described item in a statement template.
owl:minQualifiedCardinality	Allowed minimum occurrence of the used property in a statement template.
owl:maxQualifiedCardinality	Allowed maximum occurrence of the used property in a statement template.
rdfs:comment	To describe value class or value range of the used property, vocabulary encoding scheme and syntax encoding scheme of the property value.

Figure 5.5 shows a part of DSP of DPLA MAP V4 in RDF Turtle syntax, where a resource *<http://DSP/V4/Collection>* is a DT, which is an instance of the class *"dsp:DescriptionTemplate"*. This instance of DT *<http://DSP/V4/Collection>* has two STs, which are identified by *<http://DSP/V4/Collection/CollectionTitle>* and *<http://DSP/V4/Collection/CollectionDescription>*, and these STs are instances of the class *"dsp:StatementTemplate"*.

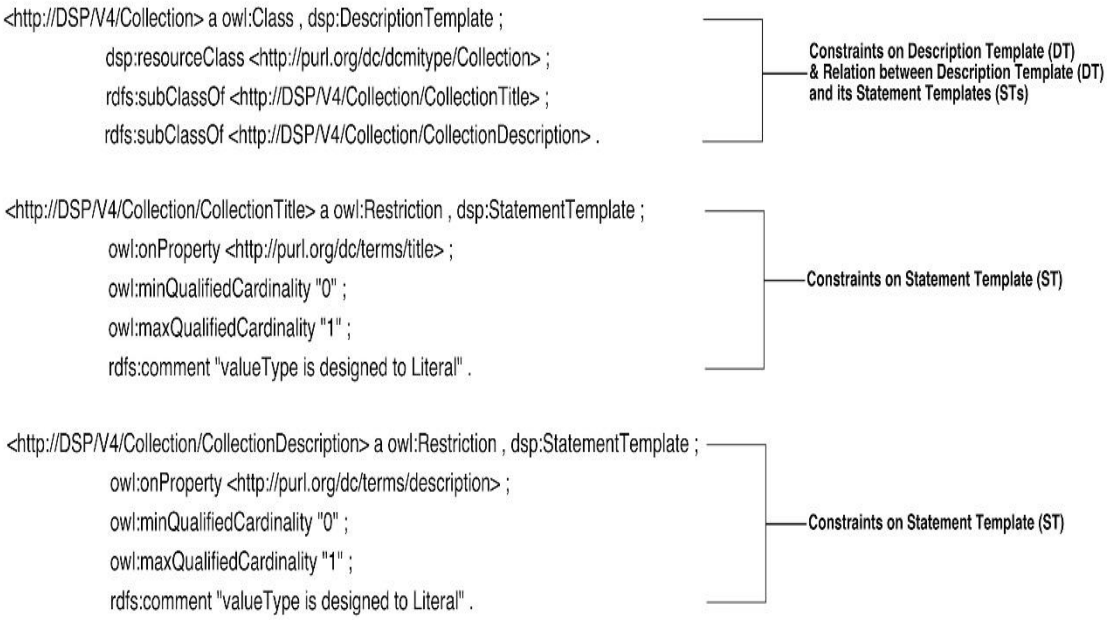

Figure 5.5: Partial RDF data of Description Set Profile of DPLA MAP V4.

5.3.3. Generation of DSP-PROV Provenance Description of DPLA MAP

Figure 5.6 shows generation process of formal provenance description of DPLA MAP. The author uploaded previously created DSP RDF data of DPLA MAP into a Virtuoso RDF Store. The author developed a program using a Ruby implementation of a SPARQL client for pure-Ruby library RDF.rb to work with the RDF data. The developed program enables the following functions: extracting data through a SPARQL Endpoint; comparing the extracted data from two neighboring versions of DSP of DPLA MAP for tracking the structural changes of DPLA MAP; identification of the deleted, added and derived structural schema instances and creation of formal provenance description in Turtle serialization syntax.

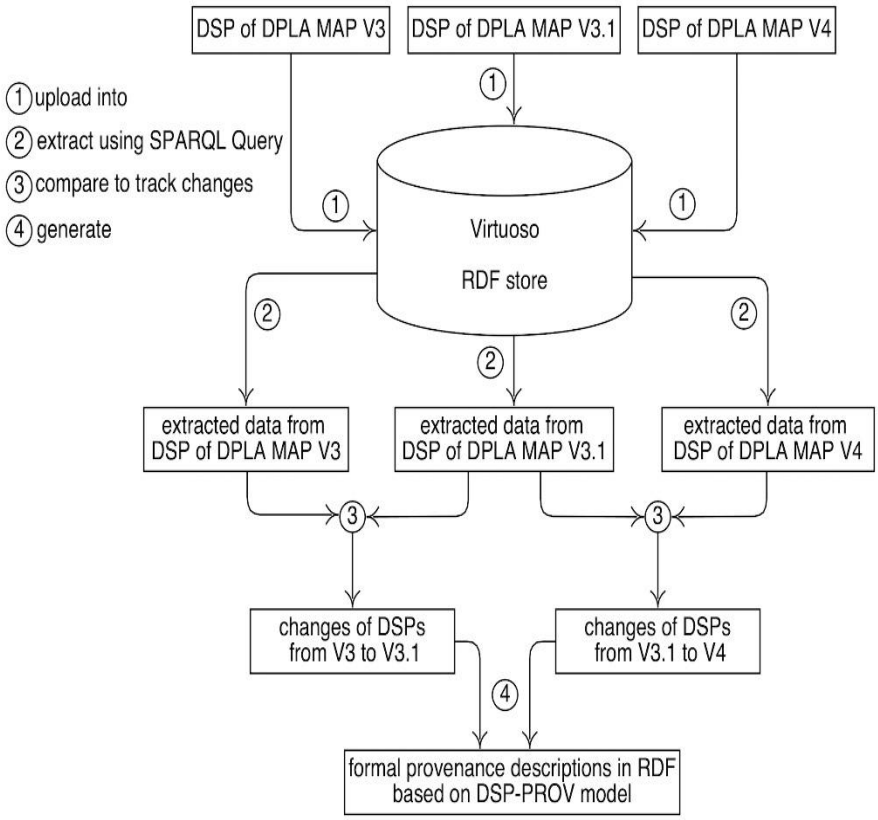

Figure 5.6: Generation process of formal provenance description using DSP-PROV model.

5.3.4. RDF Models for Creation of Formal Provenance Description of Metadata Application Profile

Figure 5.7 shows RDF graphs to create provenance description of MAPs in the following three patterns. (a) Deletion: The deleted structural schema instance was invalidated by its influencing Deletion Activity. (b) Addition: The added structural schema instance was generated by its influencing Addition Activity. (c) Revision: The structural schema instance in the subsequent version was derived from its corresponding structural schema instance in the previous version and was generated by its influencing Revision Activity. Structural schema instance defined in the previous version was used and was invalidated by the influencing Revision Activity.

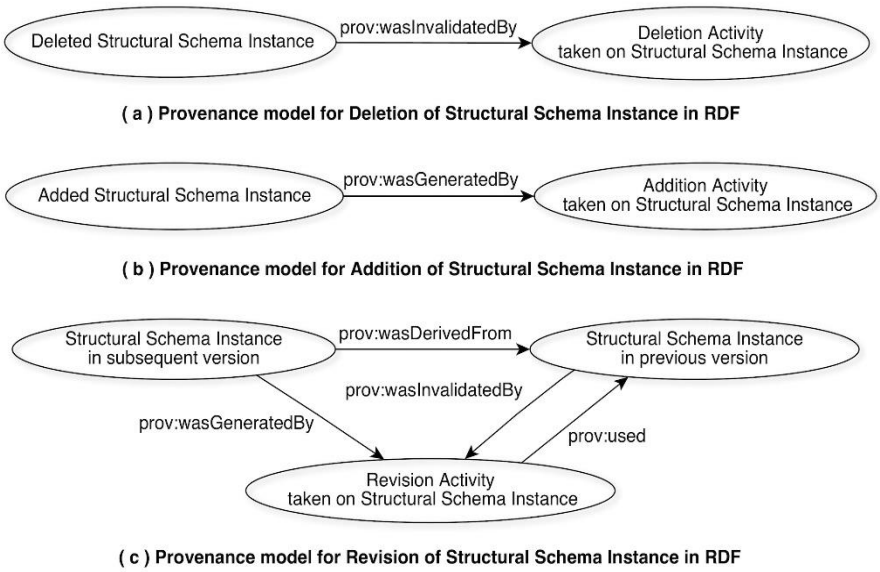

Figure 5.7: Provenance model for deletion/addition/revision of structural schema instance.

Figure 5.8 shows a model in RDF graphs to create provenance descriptions that describe the relationships among Activities. As illustrated in Figure 5.8, a Revision Activity acted upon containing Entity (e.g., a Description Template) is connected with its sub-activities (i.e., Deletion, Addition, Revision) acted upon its contained Entity (e.g., a Statement Template of a Description Template) via the property *"dcterms:hasPart"*.

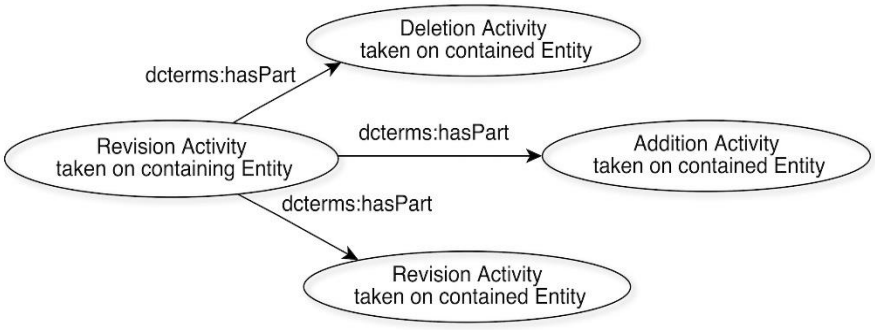

Figure 5.8: Provenance model for provenance descriptions among Activities.

RDF graphs in Figure 5.9 describe the following change: property to describe "Collection Title" in Class *"dcmitype:Collection"* is changed from *"dc:title"* in DPLA MAP V3.1 to *"dcterms:title"* in DPLA MAP V4. These triples are created following RDF models in Figure 5.7(c) and Figure 5.8. Figure 5.9 shown below provides formal provenance description that reveals the following changes in the different levels.

(1) Description Template (DT) level: Revision of the DT that defines the constraints to describe “Collection”.

(2) Statement Template (ST) level: Revision of the ST that defines the constraints to describe “Collection Title” in the above DT.

(3) Structural Constraint (SC) level: Revision of the SC that defines the constraint of used property in the above ST.

The prefix *“dspprov”* is the namespace for the classified activities. The property *“rdf:type”* describes the class that an instance belongs to. As shown in Figure 5.9, an Revision Activity instance of “dspprov:RevisionOnDT”, *“dspprov:RevisionOnST”*, and *“dspprov:RevisionOnSC”*, respectively occurred in the above change case.

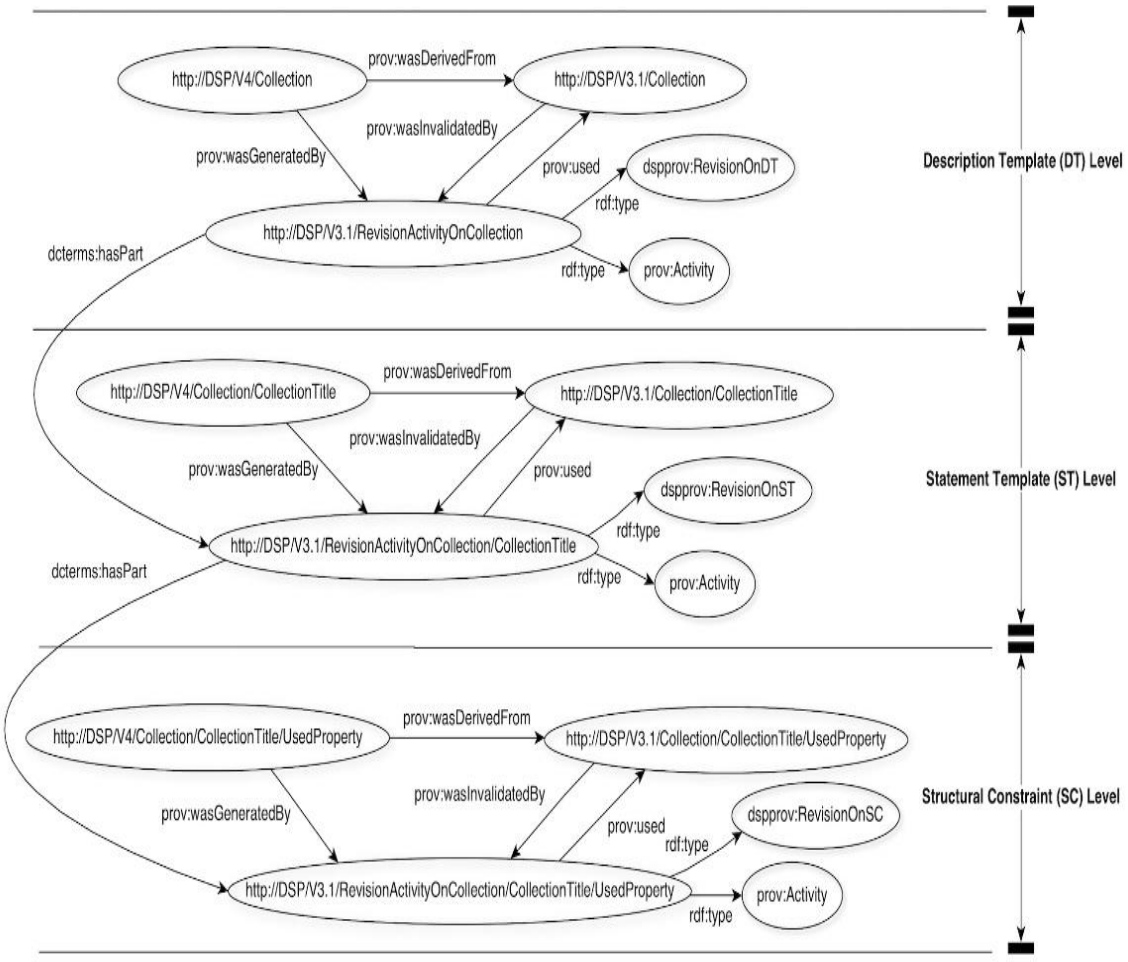

Figure 5.9: Example of formal provenance description based on DSP-PROV model.

5.4. Evaluation of DSP-PROV Model

5.4.1. Correctness Check of Semi-formal and Formal Provenance Description of DPLA MAP

DSP-PROV is evaluated by checking provenance descriptions based on DSP-PROV and a set of change logs of DPLA MAP. The former is given in a formal syntax defined in RDF and the latter, which is called semi-formal description, is written in English with a structured description format. Figure 5.10 briefly gives the overall image of the evaluation process, which refers to the following three datasets and specific procedures.

Dataset 1: Change Logs of DPLA MAP in PDF files. Changes from DPLA MAP V3 to V3.1 and from DPLA MAP V3.1 to V4 are described in change logs of DPLA MAP. The changes are manually recorded in English and in a structure defined for the change log description using controlled phrases. Thus, information about changes from a version of DPLA MAP to its next version is semi-formally presented for avoiding misunderstanding. Some examples of the statements in the change logs are shown in Table 5.5.

Dataset 2: DSP of DPLA MAP in RDF. The author created DSP RDF data of the three versions of DPLA MAP in RDF as shown above. These three versions of DSP RDF data are uploaded into a Virtuoso RDF store and accessible through a SPARQL endpoint.

Dataset 3: DSP-PROV provenance descriptions in RDF. The author applied the proposed model to the provenance descriptions created from the DPLA MAP and stored the formal provenance description in DSP-PROV into a Virtuoso RDF store.

The whole evaluation process includes two parts, i.e., Evaluation 1 and 2 shown below.

Evaluation 1: check if there are errors in the change log statements in Dataset 1. The author examined the change log statements with DSP of DPLA MAP in Dataset 2.

Evaluation 2: check if there are errors in the DSP-PROV descriptions in Dataset 3. The author examined the DSP-PROV descriptions with DSP of DPLA MAP in Dataset 2.

The following paragraphs explain the whole evaluation process.

Step 1: identification of changed classes and properties in change logs (Dataset 1). The author manually examined every single change statement in the change logs and identified the revisions of changed classes and properties described in the logs.

Step 2: trace of definitions of changed classes and properties in the DSP RDF data (Dataset 2). The author wrote SPARQL queries to track definitions of the changed classes and properties from the two consecutive versions of DSP of DPLA MAP. The query results show the differences between the consecutive versions.

- If a new class or a new property appears in a following version, it means an addition of a structural schema instance.
- If an existing class or an existing property disappears in a following version, it means a deletion of a structural schema instance.
- If structural constraints of a class or a property between consecutive versions differ, it means a revision of a structural schema instance.

Step 3-A (Evaluation 1): correctness check of change logs (Dataset 1). The author checked the correctness of change logs by comparison between the change logs and the differences gained from Step 2. The criteria below are used to identify errors.

- Criteria 1 Naming Convention Violation: The first letter of a class name is not capitalized, the first letter of a namespace is capitalized, the first letter of the first word of a property name is not in lowercase, or the first letter of the words except first word of a property name is not in uppercase
- Criteria 2 Name Incorrectness: Wrong property names and class names
- Criteria 3 Namespace Inconsistency: Wrong namespace names and URIs of class and property (including no recording of namespace of class and property)
- Criteria 4 Structural Inconsistency: Wrong structural constraints
- Criteria 5 Duplicated Change Logs: Two or more descriptions about a single change
- Criteria 6 Missing Records: Changes that are not recorded

Step 3-B (Evaluation 2): correctness check of the DSP-PROV descriptions (Dataset 3). The author checked correctness of the formal DSP-PROV descriptions by comparing them with the differences gained from Step 2. This check was conducted as a series of SPARQL queries to find inconsistency between the Activities and Entities.

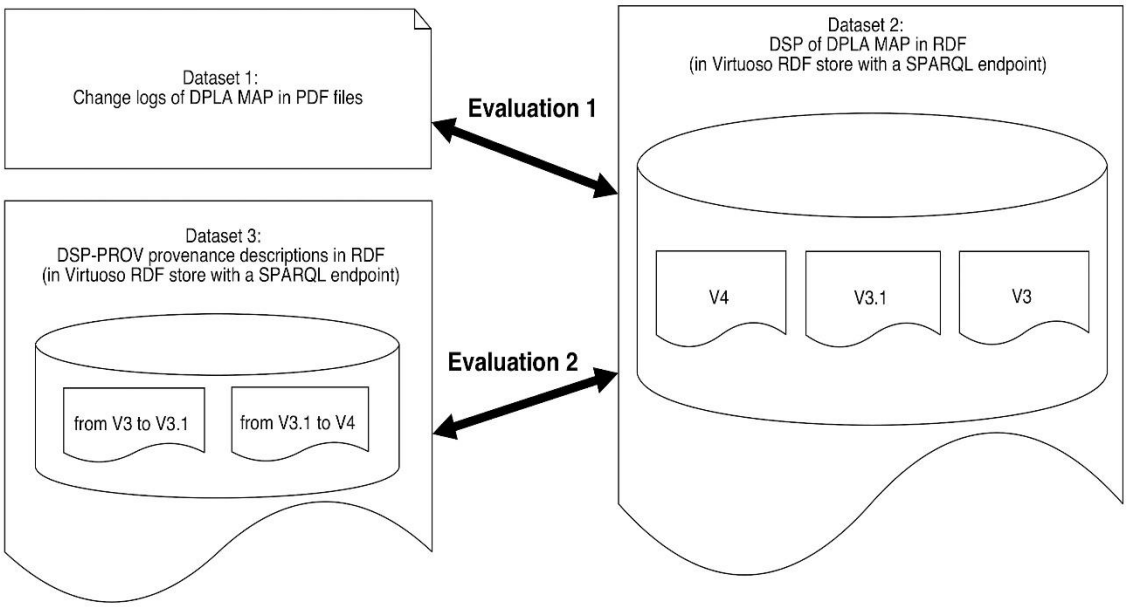

Figure 5.10: Correctness check of provenance descriptions of DPLA MAP.

5.4.2. Errors Found in Semi-formal Provenance Description of DPLA MAP

Table 5.5 shows the change logs and their corresponding errors identified by the above procedures for Evaluation 1. Table 5.5 (a) and (b) show the change log statements and errors found between versions 3 and 3.1 and errors found between versions 3.1 and 4, respectively. Criteria and Errors columns in Table 5.5 show the specific criteria classified in the above and reasons of errors found for changes in Change Log Statements. Appearance Position of Changes columns show the changes appearance in the two consecutive versions. N/A means “not applicable”.

Table 5.5: Errors in change logs of DPLA MAP.

a. Errors found in the change logs from version 3 to 3.1

Criteria	Change Log Statements	Appearance Positions of Changes		Errors
		Version 3	Version 3.1	
2	▪ Added "Genre" property to dpla:SourceResource class	N/A	Page 5	"Genre" is the label rather than the name of the referred property, whose name should be edm:hasType.
3	▪ Changed obligations for "Collection Title" and "Collection Description" in dc:Collection class	Page 4	Page 6	The namespace of "dc:Collection class" is incorrectly recorded, which should be dcmitype:Collection.
3 and 4	▪Range of dc:date and dc:temporal on dpla:SourceResource class set to edm:TimeSpan	Page 3, Page 4	Page 4, Page 5	The namespace of "dc:temporal" is incorrectly recorded, which should be dcterms:temporal; There is no range set to dcterms:temporal both in V3 and V3.1.
2 and 5	▪ Added "Standardized Rights Statement" property to edm:WebResource class	N/A	Page 8	"Standardized Rights Statement" is the label rather than the name of the referred property, whose name should be edm:rights; This change is recorded twice.

b. Errors found in the change logs from version 3.1 to 4

Criteria	Change Log Statements	Appearance Positions of Changes		Errors
		Version 3.1	Version 4	
1, 2 and 3	・Change from literal to Ref value for the following properties in dpla:sourceResource ・creator ・genre ・language ・place ・publisher ・relation ・subject ・type	Page 4, Page 5	Page 5, Page 6, Page 7, Page 8	The first letter in class name "dpla:sourceResource" is not capitalized, which should be dpla:SourceResource; The property for "genre" is incorrectly recorded, which should be edm:hasType; The properties without namespaces in the change statement are not completely recorded.
1, 2 and 3	・ Addition of the following properties to dpla:sourceResource ・alternative title ・replacedBy ・replaces ・rightsholder	N/A	Page 5, Page 7	The first letter in class name "dpla:sourceResource" is not capitalized, which should be dpla:SourceResource; The property for "alternative title" is not correctly recorded, which should be dcterms:alternative; The properties without namespaces are not completely recorded, which should be dpla:isReplacedBy, dpla:replaces, dcterms:rightsholder.
1 and 6	・Change from DC to DCTERMS namespaces for the following properties in dpla:sourceResource ・contributor ・creator ・description ・identifier ・language ・subject ・title ・type	Page 4, Page 5	Page 5, Page 6, Page 8	The first letter in class name "dpla:sourceResource" is not capitalized, which should be dpla:SourceResource; There is no recording about "publisher", whose namespace is also changed from dc to dcterms.
1 and 3	・ Creation of Agent class to describe persons or organizations referred to in dpla:sourceResource and ore:Aggregation	N/A	Page 5, Page 7, Page 9, Page 10, Page 11	The first letter in class name "dpla:sourceResource" is not capitalized, which should be dpla:SourceResource; The namespace of the "Agent class" is not recorded, which should be edm:Agent.
3	・Creation of Concept class for the description of any topic or subject heading	N/A	Page 13, Page 14	The namespace of the "Concept class" is not recorded, which should be skos:Concept.

(continued)

1 and 6	・Skos:Concept for range of subject	N/A	Page 8	The first letter in namespace of skos:Concept class is capitalized; The range addition to property dcterms:subject is not completely recorded.
1 and 2	・Deprecation of State Located In property within dpla:sourceResource	Page 5	N/A	The first letter in class name "dpla:sourceResource" is not capitalized, which should be dpla:SourceResource; "State Located In" is the label rather than the name of the referred property, whose name should be edm:currentLocation.
1 and 2	・Change to Ref pointer for Digital Resource Source Record instead of storing original record as a literal within dpla:sourceResource	Page 9	Page 10	The first letter in class name "dpla:sourceResource" is not capitalized, which should be dpla:SourceResource; The class name "dpla:sourceResource" is incorrectly recorded, which should be ore:Aggregation.
1 and 2	・Addition of the Preview property for thumbnail pointers within edm:webResource	N/A	Page 11	The class name "edm:webResource" is incorrectly recorded, which should be ore:Aggregation; The property name is not correctly recorded, which should be edm:preview.
2 and 3	・Addition of the following properties in the edm:Place class ・Latitude ・Longitude ・Altitude ・Geometry ・Parent Feature	N/A	Page 14, Page 15	The namespaces of the properties are not recorded; These are labels rather than the names of the referred properties, which should be wgs84:lat, wgs84:long, wgs84:alt, geojson:geometry, gn:parentFeature.
2, 4 and 6	・Addition of the following properties in the edm:Place class ・Parent country ・Same as	N/A	N/A	There is no definition of "Parent country" in V4 and the added property should be gn:countryCode; There is no definition that relates to "Same as" in V4.
2 and 3	・Deprecation of the following properties in the edm:Place class ・City ・State ・County ・Region ・Country ・Coordinates	Page 6, Page 7	N/A	These are labels rather than the names of the referred properties; The namespaces of the properties are not recorded, which should be dpla:city, dpla:state, dpla:county, dpla:region, dpla:country, wgs84:lat_long within the class dpla:Place in version 3.1.
6		N/A	Page 15, Page 16	The following changes were not recorded: addition of properties skos:note, skos:inScheme, skos:exactMatch, and skos:closeMatch in the edm:Place class.

This paragraph explains the three steps in Evaluation 1 using a change example from V3.1 to V4 of DPLA MAP. Errors in the change log statement "Addition of the Preview property for thumbnail pointers within edm:webResource" in Table 5.5 was found by the following steps. The author first identified the changed property ("Preview property") and changed class (*"edm:webResource"*). Then, the author wrote SPARQL queries 1 and 2 shown in Table 5.6 to track constraints related to "Preview" from Description Set Profiles of DPLA MAP V3.1 and V4, respectively. Next, the author compared and analyzed their query results of definitions related to "Preview" in the two consecutive versions, from which errors in the change log statement were identified based on the above criteria.

Table 5.6: SPARQL queries for checking correctness of provenance description.

Query 1 for Description Set Profile of DPLA MAP V3.1	Query 2 for Description Set Profile of DPLA MAP V4
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX dsp: <http://purl.org/metainfo/terms/dsp#> *SELECT (COUNT(?ST) AS ?countV3_1)* FROM <http://DPLA/MAP/DSP/V3.1> WHERE { ?DT a dsp:DescriptionTemplate; dsp:resourceClass ?resourceClass; rdfs:subClassOf ?ST. *FILTER regex(?ST,"Preview")* }	PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX dsp: <http://purl.org/metainfo/terms/dsp#> *SELECT (COUNT(?ST) AS ?countV4) ?ST ?usedProperty ?DT ?resourceClass* FROM <http://DPLA/MAP/DSP/V4> WHERE { ?DT a dsp:DescriptionTemplate; dsp:resourceClass ?resourceClass; rdfs:subClassOf ?ST. ?ST owl:onProperty ?usedProperty. *FILTER regex(?ST,"Preview")* }

The values of variables count V3_1 in Query 1 and count V4 in Query 2 respectively return the occurrence number of statement template related to "Preview" in Description Set Profile V3.1 and V4. Query 1 and Query 2 results indicate no statement template and one statement template defining constraints on "Preview", respectively. This means, a new statement template defining constraints on "Preview" was added from Description Set Profile V3.1 to V4. Therefore, "addition" itself is correctly recorded in the change log statement between the two requested versions. The exact information about "what is added" can be known by Query 2 result as explained below.

Query 2 result indicates structural constraints of the newly added statement template identified by URI *<http://DSP/V4/Aggregation/Preview>*, which is one component of the description template identified by URI *<http://DSP/V4/Aggregation/>*. The exact name of the added property with prefix of namespace is *"edm:preview"*, whose full URI is *<http://www.europeana.eu/schemas/edm/preview>*. Therefore, the property name "Preview" in the change log is incorrect in accordance with Criteria 2. The newly added property *"edm:preview"* is used to describe class *"ore:Aggregation"* with full URI *<http://www.openarchives.org/ore/terms/Aggregation>*. According to Criteria 1 and 2, a

conclusion was achieved that the newly added property *"edm:preview"* in the scope of class "edm:webResource" is incorrectly recorded in the above change log statement.

The author also tracked in DSP-PROV provenance description corresponding to the above change case, which describes addition of the Statement Template (ST) <*http://DSP/V4/Aggregation/Preview*> defining constraints on "Preview". The ST <*http://DSP/V4/Aggregation/Preview*> was created by an activity instance <*http://DSP/V3.1/AdditionActivityOnAggregation/Preview*> of class *"dspprov:AdditionOnST"*. The following triples serialized in Turtle syntax is consistent with what were gained from Queries 1 and 2. In the Turtle syntax, the token *"a"* stands for *"rdf:type"*, which is used to assert an instance of a class.

```
<http://DSP/V4/Aggregation/Preview> a prov:Entity;
                                    prov:wasGeneratedBy <http://DSP/V3.1/AdditionActivityOnAggregation/Preview>.
<http://DSP/V3.1/AdditionActivityOnAggregation/Preview> a prov:Activity, dspprov:AdditionOnST.
```

Besides missing records in Table 5.5, the comparison conducted between the semi-formal and formal provenance description of DPLA MAP shows other several changes missing in the change logs of DPLA MAP. For instance,

- From the version 3 to version V3.1, there is no recording of addition of range *"edm:ProvidedCHO"* to property *"edm:aggregatedCHO"* and addition of range *"edm:WebResource"* to property *"edm:object"* and *"edm:isShownAt"*;
- From the version V3.1 to version V4, there is no recording of addition of range *"skos:Concept"* to property *"edm:hasType"* and addition of range *"edm:TimeSpan"* to property *"dcterms:temporal"*.

5.4.3. Advantages of Formal Provenance Description of DPLA MAP

In the case of semi-formal provenance description of DPLA MAP, the connection between a version of metadata application profile with its next version has to be manually traced. However, in the case of formal provenance description, the relationships between the two consecutive versions of a metadata application profile can be automatically traced.

RDF graphs in Figure 5.11 provide DSP-PROV provenance descriptions, which correspond to semi-formal provenance description "Added dpla:intermediateProvider to ore:Aggregation" in the change log of DPLA MAP from V3 to V3.1. The RDF graphs in the upper rectangle describe a Revision of a DT that defines constraints on class *"ore:Aggregation"*. The Description Template (DT) identified by URI <*http://DSP/V3.1/Aggregation*> in DPLA MAP V3.1 is derived from the

DT identified by URI <*http://DSP/V3/Aggregation*> in DPLA MAP V3. The Revision Activity identified by URI <*http://DSP/V3/RevisionActivityOnAgrregation*> caused invalidation of the DT identified by URI <*http://DSP/V3/Aggregation*> and generation of the DT identified by URI <*http://DSP/V3.1/Aggregation*>. The two DTs in the two consecutive versions of DPLA MAP formally connect with each other through their derivation relationship, which can be automatically traced. The RDF graphs in the lower rectangle show Addition of a Statement Template (ST) that defines constraints on property *"dpla:intermediateProvider"*. The Addition Activity in the lower rectangle generated the ST identified by URI <*http://DSP/V3.1/Aggregation/IntermediateProvider*>. The Revision Activity in the upper rectangle links to its sub-activity, i.e., the Addition Activity in the lower rectangle through property *"dcterms:hasPart"*.

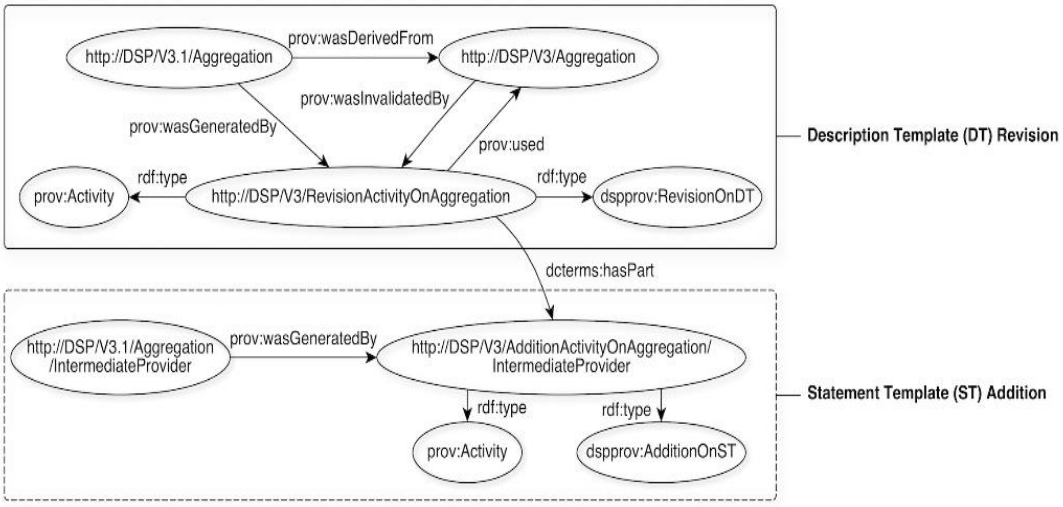

Figure 5.11: Example of formal provenance description of DPLA MAP.

5.5. Summary

This chapter presents the development of DSP-PROV model for tracking the changes of structural features of metadata schemas based on standards that are widely accepted on the Web, i.e., Singapore Framework for Dublin Core Application Profile and W3C PROV standard. To propose a formal model for describing provenance of metadata schemas, the author firstly collected several metadata schemas along with their revision history. Based on analysis of their revision history and usage of the collected resources, this study adopted a model-based approach to define the DSP-PROV model for provenance description of metadata application profiles.

The author applied DSP-PROV to a set of documents of metadata schema and revision history provided by DPLA to show advantages of DSP-PROV. DPLA MAP was carefully selected as a case study among several metadata schemas based on the criteria defined in this study. The DSP-PROV model was evaluated through the case study of DPLA MAP to compare formal provenance description in DSP-PROV and semi-formal change log description in English. This chapter shows advantage of formal provenance description based on the proposed DSP-PROV model to semi-formal provenance description in a natural language.

6. Provenance for Long-term Maintenance of Metadata Vocabulary

It is widely known that term definitions and term usage change over time. Metadata schemas are also a digital object shared in the Linked Data environment as well. Metadata vocabularies are the semantic basis for metadata sharing across communities and over time. Collecting and maintaining metadata vocabularies is a fundamental task for memory institutions to keep their memory materials consistent regardless of the genres, types and formats – tangible/intangible, digital/non-digital. Even if the meanings of terms would be changed little by little the semantic change over time may be large. Therefore, keeping the changes traceable by machines is important to keep metadata consistently interpretable. Metadata schemas for applications in the networked information environment are developed using one or more standards. “Mixing and Matching” and “Do not re-invent wheels” which are the motto of DCMI tell us that we should re-use existing standards and combine them in accordance with the requirements given to specific applications. This means that the semantic definitions of old metadata terms should be maintained, and the change histories of the terms should be traceable.

This chapter focuses on consistent maintenance of metadata vocabularies and metadata terms. This is because the changes of definitions of a metadata term may not always be recorded appropriately. The definition of a metadata term may include meaning and usage of the term, relationships to other terms, human-readable labels, and so forth. Metadata terms are usually defined as a set of terms, which is called a metadata vocabulary. The author aims to propose a metadata model designed to keep track of the changes to definitions of metadata terms and metadata vocabularies in this chapter.

6.1. Features of Metadata Vocabulary

Metadata Vocabulary and Terms In the library community, commonly used metadata vocabularies are controlled vocabularies and metadata element sets (Hyland et al., 2013; Isaac et al., 2011), e.g., subject headings, authority files, Resource Description and Access (RDA) element sets, and RDA value vocabularies. A metadata vocabulary is a set of metadata terms. Here, “metadata vocabulary” is used as a generic concept that includes two types, i.e., property vocabulary and value vocabulary. A property vocabulary is a set of terms expressing attributes of a resource and relationships between resources, which is often called metadata element set, e.g., Dublin Core metadata element set and BIBFRAME vocabulary. A value vocabulary is a set of terms expressing classes of resources and encoding schemes of property values, e.g., Library of Congress Subject Headings (LCSH). To propose general provenance description model for tracking

primitive changes of metadata terms in metadata vocabularies, this study defines "Term" and "Term Definition" as follows.

Term in a metadata vocabulary is an individual entity, which represents a concept, a property, a class, and a metadata vocabulary. For example, a subject heading in LCSH, property *"dct:title"*, class *"dct:Agent"*, and vocabulary encoding scheme LCSH are examples of terms. This study uses "Term" in both meanings of property vocabulary term and value vocabulary term.

Term Definition is a set of descriptions that defines features of the term. The features are the human-readable label(s) of the term, the meaning of the term, relationships between terms, usage of the term, and other information. Term Definition may be seen as a set of statements, each of which defines a feature of the term. For instance, "the broader term of Vehicles in LCSH is Transportation" is a Term Definition of Term "Vehicles"; "the label of term subject in Dublin Core metadata element set is Subject" is a Term Definition of Term *"dc:subject"*. The two examples of Term Definition can be respectively represented as RDF triples, *lcsh:sh85142531 skos:broader lcsh:sh85137027* and *dc:subject rdfs:label "Subject"@en*. The *lcsh:sh85142531* stands for "Vehicles" while the *lcsh:sh85137027* stands for "Transportation".

Metadata vocabularies should be maintained to keep metadata terms consistently interpretable. The definition of a metadata term may be changed, e.g., renaming of a term, revision of the meaning of the term, and revision of relationships to other related terms. It is crucial to trace changes of metadata terms in metadata vocabularies. Provenance description for long-term maintenance of metadata vocabularies is primarily the series of activities that have taken place on metadata vocabularies and their terms. The author proposed a model to describe provenance description of metadata vocabularies based on W3C PROV. Entities and activities based on the relations defined in W3C PROV are classified to describe primitive changes of metadata terms in metadata vocabularies. The recorded entities and activities are traceable to provide evidence for change tracking, which brings the benefits of provenance description of metadata vocabularies, e.g., preventing misinterpretation and auditing inconsistencies of metadata vocabularies. These benefits are valuable for the long-term maintenance of metadata vocabularies throughout their life cycle.

Provenance of metadata vocabularies is a record that describes the agents, activities, and entities involved in the lifecycle of metadata vocabularies. Provenance of metadata vocabularies includes information about how metadata terms in a metadata vocabulary and its term definitions come to a specific state. The definitions of metadata terms can change over time. For instance, a term can be split into two related terms, or the semantic relationship between two terms can change over time. Those who are responsible for maintaining metadata vocabularies need to pay attention to the changes and documentation of the changes.

6.2. Primitive Changes of Metadata Vocabulary

Entities and Activities for Provenance Description of Metadata Vocabularies. Vocabulary, Term, and Term Definition are classified as three subtypes of PROV Entity to describe provenance of metadata vocabularies. As illustrated above, a Term can be a concept or a class or a property. In the case of a concept, its definition may include its narrower term(s), broader term(s), association/related term(s), and other information. In the case of a class, its definition may include a description of its meaning, a label(s), a URI, super-class(es), sub-class(es), used property(ies), and other information. In the case of a property, its definition may include a description of its meaning, a label(s), a URI, super-property(ies), sub-property(ies), domain, range, expected value, and other information.

To describe the provenance of metadata vocabularies, Activities acting on the previously classified Entities are categorized into the following types, i.e., Revision, Addition, Deletion, and Replacement. Table 6.1 shows the correspondence of the classified Activities to the classified Entities. The mark "○" means "applicable" and "×" means "not-applicable". Table 6.2 illustrates the classified Activities with their names and definitions. It is notable that replacement of term can be the following cases, e.g., a composite term was split into more than one term; or more than one term was merged to a term; or a term was replaced by another term. Table 6.3 provides change types of metadata vocabularies as well as their terms with specific examples, which are mainly from the changes between BIBFRAME 2.0 vocabulary (BIBFRAME 2.0 vocabulary list view, 2016) and BIBFRAME 1.0 vocabulary (BIBFRAME 2.0 specifications notes, 2016). The separation of a single term into two or more terms is called a split. An example of a split in a subject heading is given in Table 6.3.

Table 6.1: Activities acted on Entities for provenance of metadata vocabularies.

Subtypes of PROV Entity	Subtypes of PROV Activity			
	Revision	Addition	Deletion	Replacement
Vocabulary	○	×	×	×
Term	○	○	○	○
Term Definition	○	○	○	○

Table 6.2: Definitions of the classified Activities for provenance of metadata vocabularies.

Activity Name	Definition
RevisionOnVocabulary	The revision of the contents or information of a metadata vocabulary
RevisionOnTerm	The revision of a term of the metadata vocabulary
AdditionOnTerm	The addition of a term
DeletionOnTerm	The deletion of a term
ReplacementOnTerm	The replacement of term(s) by other term(s)
RevisionOnTermDefinition	The revision of a term definition
AdditionOnTermDefinition	The addition of a term definition
DeletionOnTermDefinition	The deletion of a term definition
ReplacementOnTermDefinition	The replacement of a term definition by another term definition

Table 6.3: Primitive change types of metadata vocabularies and their terms with examples.

Change Type	Example
Revision of a Vocabulary	BIBFRAME 1.0 vocabulary is revised to BIBFRAME 2.0 vocabulary
Revision of a Term	
Addition of a Term	Class bf:Note is newly defined in BIBFRAME 2.0 vocabulary
Deletion of a Term	Property bf:otherEditionOf that was defined in BIBFRAME 1.0 vocabulary is deleted in BIBFRAME 2.0 vocabulary
Replacement of a Term	Property bf:credits in BIBFRAME 2.0 vocabulary essentially replaces bf:creditsNote in BIBFRAME 1.0 vocabulary
Revision of a Term Definition	
Addition of a Term Definition	The inverse property to property bf:absorbed is added in BIBFRAME 2.0 vocabulary
Deletion of a Term Definition	The definitions of property bf:otherEditionOf that was defined in BIBFRAME 1.0 vocabulary is deleted in BIBFRAME 2.0 vocabulary
Replacement of a Term Definition	The expected value of property bf:copyrightRegistration is corrected in BIBFRAME 2.0 vocabulary

A revision of a vocabulary is caused by a revision of its terms. The revision of a term may be a revision of the term as an instance, or a revision of documentation of the term. For example, replacement of a single term by a set of terms is a revision of an instance, and replacement of a title text is a revision of term definition. Therefore, as shown in Figures 6.1 and 6.2, the relationships between the classified Activities are as follows. A *RevisionOnVocabulary* is comprised of *RevisionOnTerm* (zero or more than one) and *RevisionOnTermDefinition* (zero or more than one).

Given to the practical change examples of revision of a term and revision of term definitions, *RevisionOnTerm* has three general types, i.e., *AdditionOnTerm*, *DeletionOnTerm*, and *ReplacementOnTerm*; *RevisionOnTermDefinition* has three general types, i.e., *AdditionOnTermDefinition*, *DeletionOnTermDefinition*, and *ReplacementOnTermDefinition*.

The relations between Entities and Activities defined in W3C PROV include Usage, Generation, and Invalidation. Usage means utilization of an Entity by an Activity. Generation means creation of a new Entity by an Activity. Invalidation means destruction, cessation or expiry of an existing Entity by an Activity (Lebo et al., 2013). The properties *prov:used*, *prov:wasGeneratedBy*, and *prov:wasInvalidatedBy* defined in PROV-O are used to respectively describe Usage, Generation, and Invalidation. W3C PROV also defines Derivation between Entities. A Derivation is a transformation of an Entity into another, an update of an Entity resulting in a new one, or the construction of a new Entity based on a pre-existing Entity (Lebo et al., 2013). The property *prov:wasDerivedFrom* is used to directionally connect the two Entities from the new Entity to the pre-existing Entity. The overview of Vocab-PROV model for provenance description of metadata vocabularies are provided in Figure 6.2.

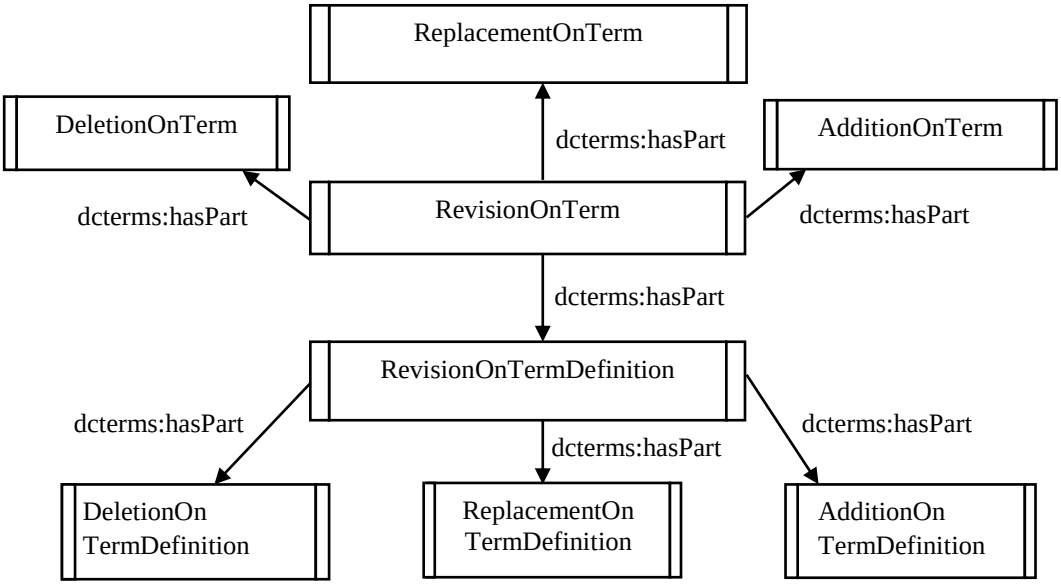

Figure 6.1: Activity relationships in the Vocab-PROV model.

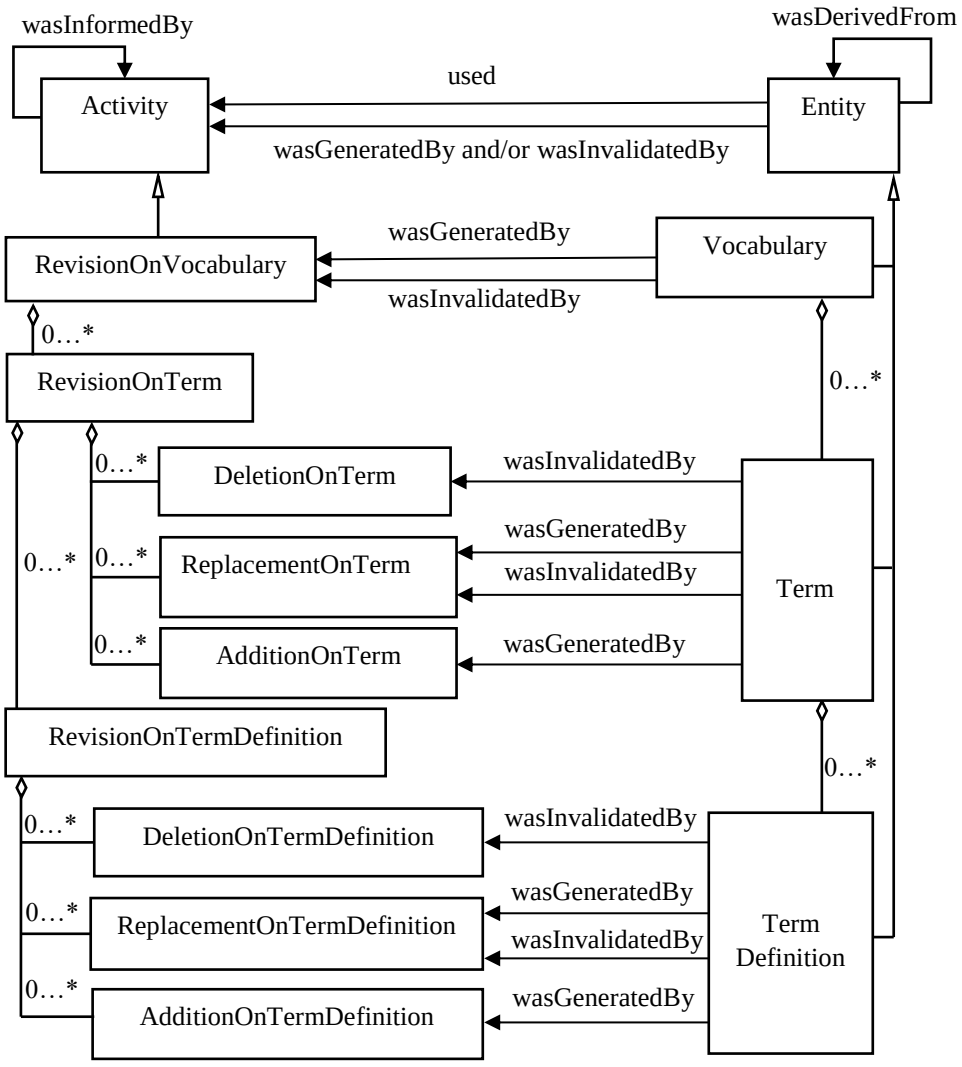

Figure 6.2: Overview of the Vocab-PROV model.

6.3. Provenance Description of Metadata Vocabulary

Figure 6.3 (a) provides provenance description in RDF graphs defined for the example of term replacement: Subject heading “Folklore, Negro” is split into “Folklore, African” and “Folklore, Afro-American”[21] (Knowlton, 2005). The classes and properties with prefix “mv” are defined in this research. The property *mv:wasSplitTo* is to describe the split of a term to more than one term. The class *mv:Term* is to assert a term of a metadata vocabulary as an instance of *mv:Term* using the property *rdf:type*. The class *mv:ReplacementOnTerm* is to assert an Activity as an instance of *mv:ReplacementOnTerm* using the property *rdf:type*. The following URIs are used to describe the

[21] This example was given by Knowlton Steven A. (2005) as already listed in the References and the contexts of the revision are out of the scope of this study.

headings: *"Folklore, Negro"* with *"http://id.loc.gov/authorities/childrensSubjects/sj96004706"*, *"Folklore, African"* with *"http://id.loc.gov/authorities/childrensSubjects/sj96004704"*, and "Folklore, Afro-American" with *"http://id.loc.gov/authorities/childrensSubjects/sj96004705"*. An Activity instance of *mv:ReplacementOnTerm* made *"Folklore, Negro"* invalidated and generated two headings, i.e., *"Folklore, African"* and *"Folklore, Afro-American"*. In the split of a LCSH term, the Library of Congress Subject Headings Supplemental Vocabularies: Children's Headings (LCSHAC) is a thesaurus that is used in conjunction with LCSH.

Figure 6.3: Example of provenance description of metadata vocabularies in RDF.

Thesaurus Entity before the split is identified by URI *"http://id.loc.gov/authorities/childrensSubjects/pv"* and the thesaurus Entity after the split is identified by URI *"http://id.loc.gov/authorities/childrensSubjects/sv"*. These thesaurus Entities are named LCSHAC PV and LCSHAC SV, respectively. Figure 6.3 (b) shows the derivation from LCSHAC PV to

LCSHAC SV. LCSHAC SV was generated by an Activity instance of *mv:RevisionOnVocabulary* and LCSHAC PV became invalidated by the same Activity instance. The class *mv:Vocabulary* is defined to assert a metadata vocabulary as an instance of *mv:Vocabulary* using the property *rdf:type*. The class *mv:RevisionOnVocabulary* is defined to assert an Activity as an instance of *mv:RevisionOnVocabulary* using the property *rdf:type*. The Activity instance of *mv:RevisionOnVocabulary* connects with the Activity instance of *mv:ReplacementOnTerm* through the property *dcterms:hasPart*, which is used to describe the inclusion relationships between Activities in this study.

6.4. Provenance Description of Semantic Change and Structural Change

A metadata application profile usually uses terms defined in existing metadata vocabularies. However, the metadata application profile may use the term meaning, which may be narrowed from the original meaning for better fit of the meaning to the application. The terms included in an existing vocabulary are usually defined within the namespace of the vocabulary without version information. Therefore, in this section, the author does not take into account the versions of the terms but focuses on the changes of term meaning defined in the metadata application profiles.

This section first provides an example of semantic change and structural change that were found from the documents of DPLA MAP. The author then discusses the provenance description about the changes using RDF graphs based on the proposed Vocab-PROV model and DSP-PROV model, respectively. Later, the relationships between the semantic change and structural change in the given change examples are briefly presented.

6.4.1. Example for Semantic Change along with Structural Change

Digital Public Library of America Metadata Application Profile (DPLA MAP) defines structural constraints of metadata, which include property, usage, obligation, range and others information in tabular form. DPLA MAP uses classes and properties from existing vocabularies, such as EDM, ORE, DC, DCTERMS, DCMITYPE, Geo vocabulary, etc. Three versions of DPLA MAP have been released, i.e., V3, V3.1, V4. DPLA MAP does not provide exact meaning and definition for its classes and properties. The value of the “Usage” column provides the kind of information related to meaning and definition of a term (i.e., class and property), which is written as the value of “Term Meaning” in the table titled “Comparison between V3.1 and V4” of Figure 6.4. DPLA MAP V3.1 provides changes from V3 to V3.1 and DPLA MAP V4 provides changes from V3.1 to V4 in a natural language.

Figure 6.4 shows change examples from DPLA MAP V3.1 to V4 that includes both structural change and semantic change. Both DPLA MAP V3.1 and V4 define property *edm:object* for class *ore:Aggregation* to describe "object". The meaning of *edm:object* in DPLA MAP V3.1 and V4 are "Unambiguous URL to the DPLA content preview" and "The URL of a suitable source object in the best resolution available on the website of the Data Provider from which *edm:preview* could be generated for use in a portal", respectively. Figures 6.4, 6.5, 6.6, and 6.7 use a short expression of the two definitions, i.e., "Unambiguous…" and "The URL…".

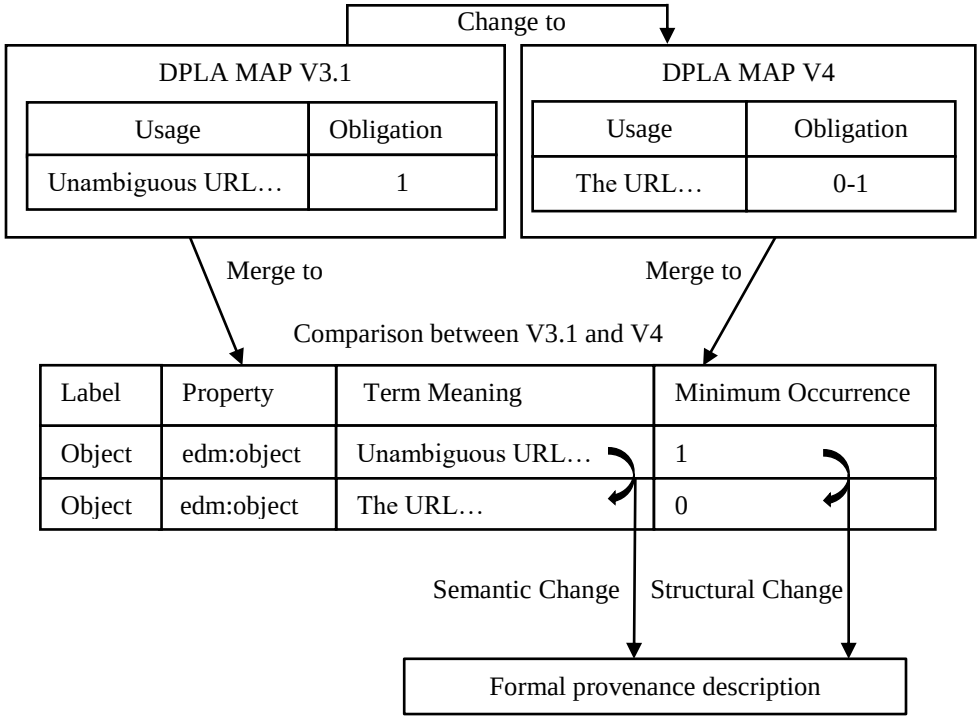

Figure 6.4: Example of semantic change along with structural change.

6.4.2. Formal Provenance Description for Semantic Change of Metadata Term

As shown in Figure 6.4, the meaning of the term *edm:object* has been changed. The change was caused by a replacement activity, which is an instance of the class *mv:ReplacementOnTermDefinition* (note: mv is the prefix for the classes of the Vocab-PROV model). According to the proposed Vocab-PROV model, the provenance description including the derivation of term definition can be formally described as the RDF graph depicted in Figure 6.5.

The meaning of the term *edm:object* is expressed as a literal value of property *skos:definition* in a rectangle (solid line). The new meaning represented in the lower dotted-rectangle was derived from the meaning represented in the upper dotted-rectangle. The newly defined meaning was

generated and the previously defined meaning became invalidated through the same Activity instance of *mv:ReplacementOnTermDefinition*.

6.4.3. Formal Provenance Description for Structural Change of Structural Constraint

As shown in Figure 6.4, the minimum occurrence of *edm:object* has been changed from "1" to "0". The change was caused by a revision activity, which is an instance of the class *dspprov:RevisionOnSC* (note: *dspprov* is the prefix for the classes of the DSP-PROV model). Figure 6.6 shows the RDF graph of provenance description about the structural constraint change.

The minimum occurrence constraint defined in the Statement Template (ST) instance that defines all the structural constraints on the property edm:object is expressed as the literal value of property *owl:minQualifiedCardinality*. The new Structural Constraint (SC) represented in the lower dotted-rectangle was derived from the previous constraint in the upper dotted-rectangle. The newly defined minimum occurrence constraint was generated and the previously defined minimum occurrence constraint became invalidated through the same Activity instance of *dspprov:RevisionOnSC*. In general, a metadata application profile uses terms that are defined in metadata vocabularies. Semantic changes of the terms used in a metadata application profile may be synchronized with structural changes of the metadata application profile. This section shows linkage of semantic change on a term and structural change in a metadata application profile.

As introduced above, Figure 6.5 shows revision of the meaning of term *"edm:object"*, which is described as the upper part of Figure 6.7. Figure 6.6 shows revision of the structural constraint on a statement template, which is described as the below part of Figure 6.7. Figures 6.5 and 6.6 show the provenance description about the semantic change and structural change in the examples given in Figure 6.4. The two statement templates defined in DPLA MAP V3.1 and V4 use properties *edm:object* in correspondence with the two consecutive versions. As shown in Figure 6.7, the connection between Figures 6.5 and 6.6 is the property constraint in the Statement Template (ST), which is expressed as the resource value of *owl:onProperty*.

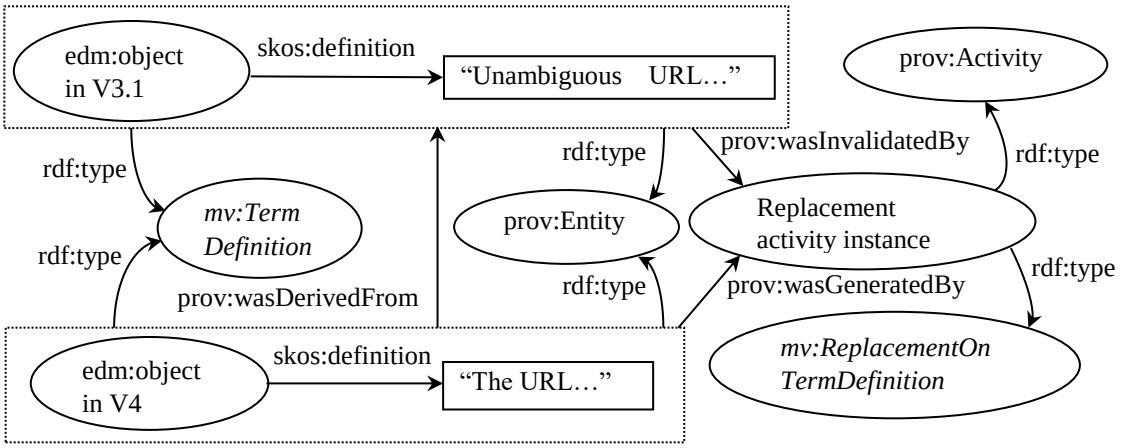

Figure 6.5: Provenance description for the above semantic change in RDF.

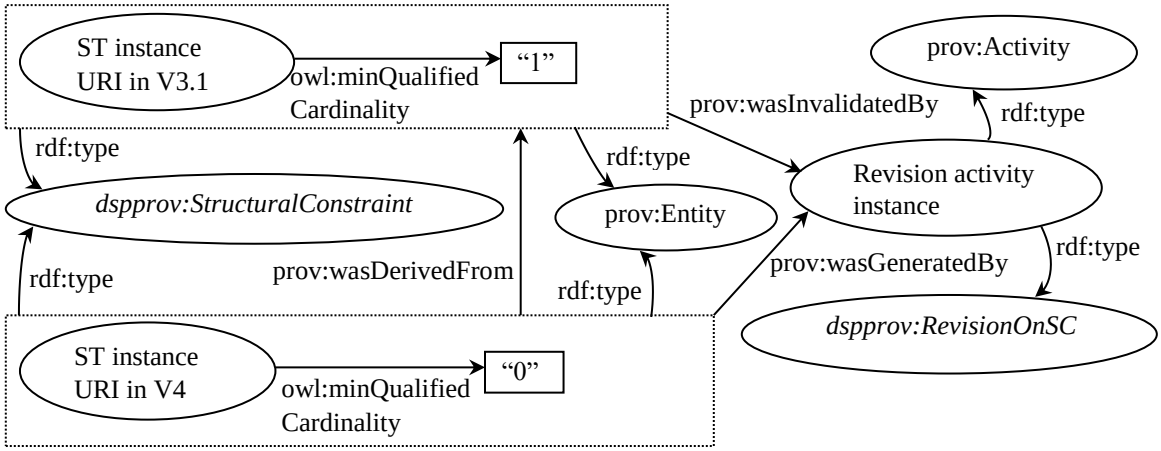

Figure 6.6: Provenance description for the above structural change in RDF.

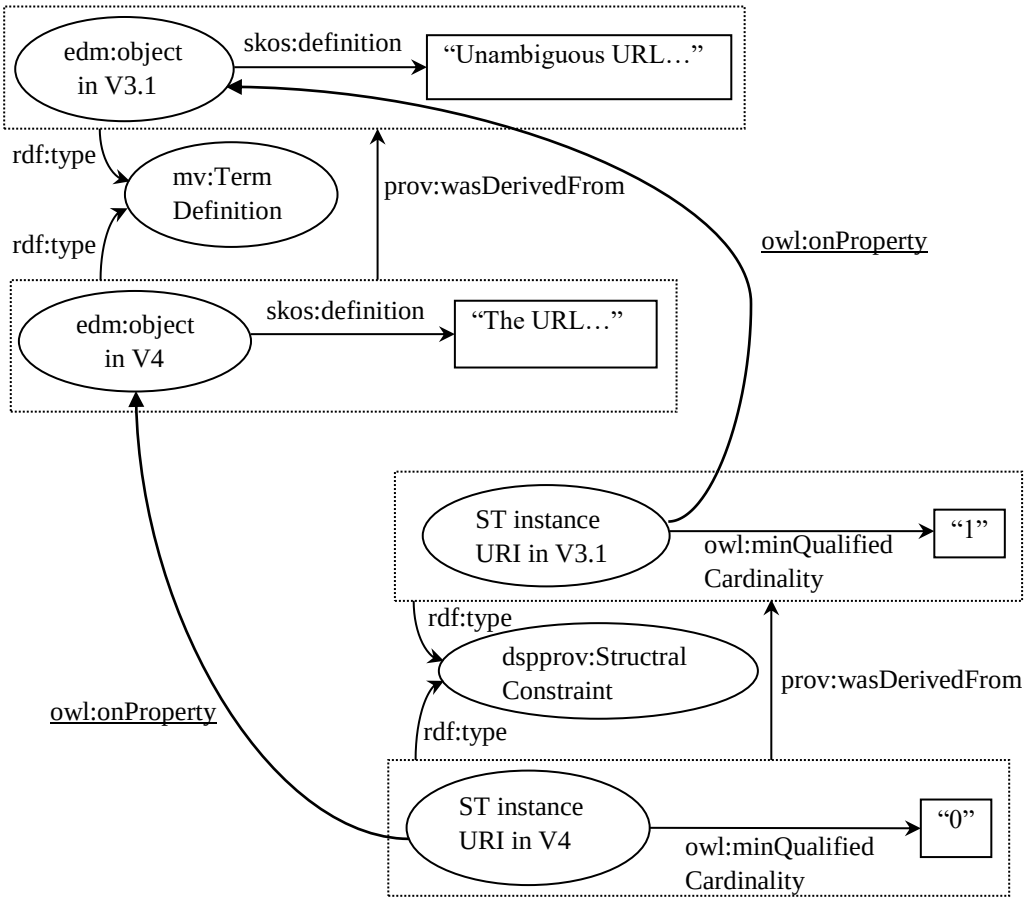

Figure 6.7: Connection between semantic change and structural change.

6.5. Summary

In this chapter, the author discussed provenance description of metadata terms and metadata vocabularies as a set of terms. The W3C PROV standard for general provenance description and RDF are adopted as the base models to formally define provenance description for metadata vocabularies. This chapter addresses the development of the Vocab-PROV model for provenance description of metadata vocabularies to track semantic changes in metadata vocabularies. Examples of formal provenance description of metadata vocabularies are provided in RDF graphs.

In this study, the author firstly examined structural changes in metadata schemas in Chapter 5. In this chapter, she examined semantic changes in metadata vocabularies. This means, structural and semantic changes in metadata schemas can be examined separately. In fact, these two kinds of changes may synchronously happen as shown in the given example in this chapter. Therefore, the author also discussed the structural constraint changes in metadata schemas along with the semantic changes in metadata vocabularies, where links can be created to connect these changes.

7. Discussion

This chapter is dedicated to a discussion of the lessons and ideas that the author has learned and gained from this study. The author discusses the limitations and implications of the proposed models as well as some open issues related to this study.

7.1. Lessons Learned from this Study

7.1.1. Metadata Preservation vs Digital Preservation

A metadata transferred on the Web is a digital object. On the other hand, a metadata is a logical data entity neutral to any physical representation as a digital object. By the nature of metadata, there is meta-metadata which is "data about metadata" and meta-meta-metadata with a meaning of "data about meta-metadata". Metadata schema is a typical meta-metadata because it is a description of metadata from the viewpoint of structural and/or semantic definition. Because of the nature of metadata, both meta-metadata and meta-meta-metadata are metadata.

Metadata instances are created as (1) a digital instance, e.g., a text file describing a book, a CSV file of bibliographic records, or (2) a logical data instance expressed as a self-contained digital object or embedded in a digital object, e.g., a metadata expressed as an RDF/XML instance and an RDFa expression embedded in an HTML document. Metadata instances are mostly but not necessarily digital objects, e.g., an XML text file and an Excel file. The technology standards for longevity of digital objects are applicable to the metadata instances.

Longevity of digital objects is well known as a crucial issue for the further progress of the networked information society. Longevity of digital objects does mean that the objects can be correctly rendered over time. However, it does not necessarily mean that future users can properly understand the content of the object. For example, a table stored in an Excel file may be rendered over time, but the attributes of the table cannot be properly understood without proper description of the meaning of the attributes and values. This table example shows a typical problem in metadata preservation. That is, metadata as a digital object may be preserved; but metadata as a semantically meaningful entity may be lost. Even if a metadata instance is encoded in XML and stored in a plain-text file, semantics of XML elements may be lost if the meanings of the tags in the XML text are not properly preserved. Thus, preservation of metadata is not same as preservation of digital objects. There are widely accepted standards for the longevity of digital objects, e.g., OAIS and PREMIS. However, there is no well-established model or standards for the longevity of metadata as a logical data entity. Therefore, the author considers and identifies keeping metadata interpretable in the future context by machines as a main goal of metadata preservation for metadata longevity.

7.1.2. Metadata Preservation Facets

This research work reveals a set of facets for the long-term maintenance of metadata – entities in different meta-levels, preservation description categories, requirements specific to metadata preservation in the LOD environment, and other aspects (Sugimoto et al., 2016). Figure 7.1 summarizes the facets described in the paragraphs below.

Facet 1: Entity Format Types – Document Files, Databases, XML Encoded Texts. Longevity management of metadata entities depends on the implementation formats of entities to be preserved. For example, it may be often the case that a metadata instance is stored in a database and an XML encoded instance is created when downloading the instance from the database. In the LOD environment, any instance that is identifiable as a resource should be given a URI. Maintaining URIs consistent is one of the key issues for metadata permanence.

Facet 2: Entity Types – Meta-Levels. As shown in Figure 7.2, there are instances of different meta-levels from level 0 to level 3. The author assumes any instances of these four categories are realized in a digital form, although they may be realized as a non-digital instance, e.g., a printed document. Instances of meta-level 1, 2 and 3 may be implemented as a document-like instance, a database record, or an XML instance encoded in a metadata description standard, e.g., RDF. Metadata preservation may be done in three approaches – document preservation, database preservation and XML encoded instance preservation in accordance with requirements at each meta-level.

Facet 3: Metadata Schema Components. This facet is for metadata schema and meta-schema entities – application profiles, metadata vocabularies for certain domains and domain-neutral standards for metadata description such as XML and namespaces. For example, Description Set Profiles and Domain Models of Singapore Framework are encoded in a formal scheme and other components are expressed as natural language texts. Preservation strategy of these components depends on the entity format types.

Facet 4: Dynamic Entities. Cases whereby metadata terms are removed from or added to a metadata schema, and when a new metadata schema is created by aggregating two existing schemas. In such cases, a mapping table are often created to map an old schema to a new schema. The mapping tables should be recorded as well as those schemas.

Facet 5: Documentation. Any document entities and activities may be recorded for use in the future. The document entities should be preserved as a part of metadata preservation. Contextual information, which may not be explicitly described in metadata schema entities, may be found in the documentation entities.

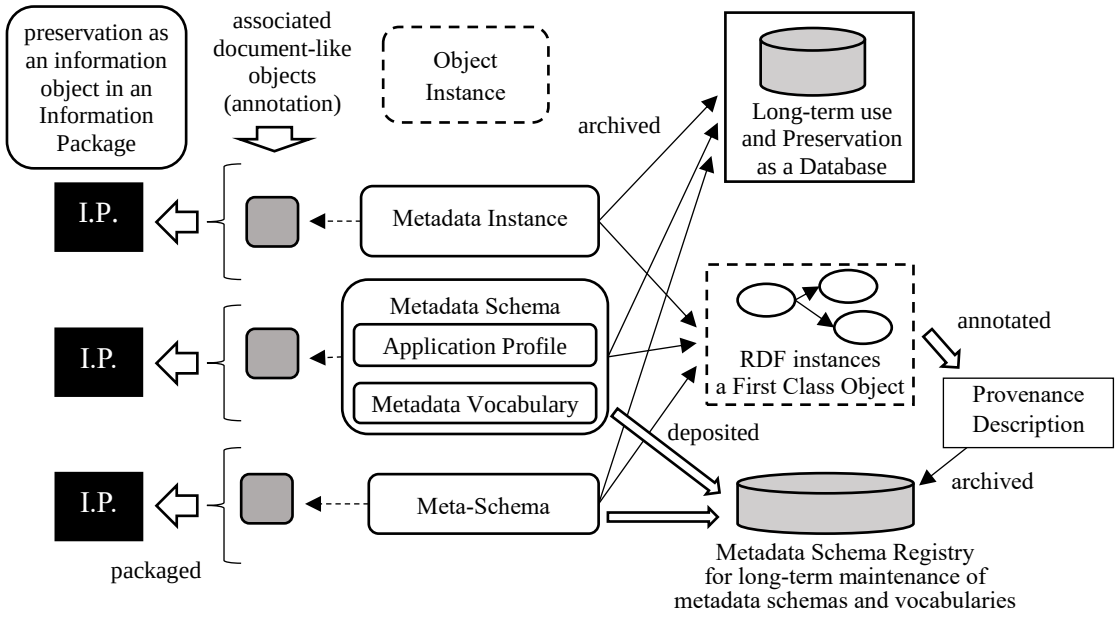

Figure 7.1: Metadata entities and preservation options.

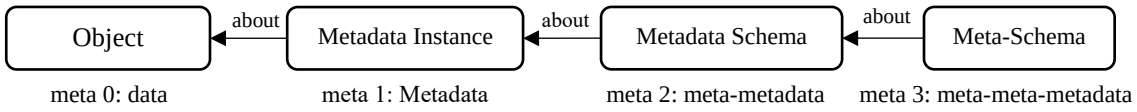

Figure 7.2: "meta-" relationships.

7.1.3. Requirements of Provenance Description on the Semantic Web

Provenance information can be described and recorded in various forms. There is a need and trend to provide provenance information as structured data with the development of Web technologies. Moreau (2010) gave a comprehensive overview of provenance in the Web context. The use of Semantic Web technologies has been advocated to facilitate provenance acquisition, representation, query and reasoning. In the Semantic Web environment, provenance description should be readable and traceable by machines, and interoperable across systems and over time.

Machine-readability: This is the minimum but fundamental requirement because provenance descriptions should be maintained and shared by machines. Standards defined for the Web have crucial roles in this aspect, e.g., RDF, PROV, OWL, etc.

Traceability: Provenance information such as where the changes come from, what activity has happened on metadata objects, what is the outcome of an activity, and what change may lead

to inconsistency, are crucial for maintenance of metadata schemas. Such provenance information should be encoded in a formal description schema for efficient and effective maintenance of schemas supported by machines and in the Web environment, e.g., RDF and SPARQL.

Interoperability: Provenance descriptions have to be interchangeable across machines and interoperable in heterogeneous system environments on the Web.

7.2. Thoughts and Ideas Gained from this Study

Providing digital collections for future use is one of the missions of memory institutions such as libraries, archives, and museums, where metadata is key for identifying and discovering their collections. Memory institutions play significant and indispensable roles in preserving and managing metadata because they need to make efforts to preserve their contents for the long-term. From this research, the author has learned that provenance is beneficial and important to metadata longevity. Moreover, there is a need to record provenance description in a structured form in the Web environment.

Publishing LOD can enhance the accessibility of the resources in the Internet. Memory institutions are keen to provide their metadata resources in the form which matches to LOD. In general, longevity is not a main issue of the LOD movement, but it is always a crucial issue for the memory institutions. Memory institutions need to make efforts to improve the accessibility of their primary resources and carry out long-term maintenance activities to keep metadata of the primary resources persistent.

As mentioned in the literature review, current provenance-related research and provenance usage in memory institutions mainly focus on authorship and ownership. Memory institutions should try to implement and extend provenance-aware services for metadata longevity. This study did not practically explore provenance services and use cases. Instead, the author briefly gave several insights of potential provenance usage in memory institutions (especially digital libraries, digital archives, and digital museums) as below. Provenance-based services can be developed to capture events and change history of digital objects and metadata objects, as well as temporal, spatial, agents and other information.

Use scenario 1: Provision of change history of subject headings using LOD technologies. One possible option can be publishing provenance of subject heading as LOD to connect the subject headings before and after the changes. It is already quite common to publish subject headings as LOD. Currently, the published subject headings as LOD lack provenance information and hence the author recommends this scenario.

Use scenario 2: Revelation of derivation of special collections (e.g., family tree, maps) using provenance graphs. Provenance tracking along with temporal or spatial information and provenance visualization (e.g., timeline) services can be considered as main functions of this scenario. As memory intuitions hold a large number of collections with temporal and/or spatial information, services integrating these temporal and spatial information along with provenance can be developed.

Use scenario 3: Tracking and management of provenance of webpages for Web archiving. In the website of an organization, there may be changes to naming and structure of the organization. Archiving not only the websites themselves but also revision history of webpages is useful for keeping institutional history for the long term. Right now, Web archiving usually harvests webpages at different time, and provenance can work as a "link" to bridge the connections between different archived versions.

7.3. Limitations and Implications of this Study

This study clarified and explored crucial issues of metadata longevity with a focus on longevity of metadata schemas and longevity of metadata vocabularies. The issues studied in this dissertation can provide reference to metadata maintainers and memory institutions. For instance, assistance in planning and strategies for metadata maintenance, metadata preservation, and risk management. However, in the LOD environment, there are other factors affecting metadata longevity, e.g., persistence of identifiers, which are just indicated without detailed discussion in this dissertation.

The author created simplified and generalized provenance models for provenance description of metadata application profiles and metadata vocabularies. The developed models enable us to track the structural changes in metadata application profiles and semantic changes in metadata vocabularies. However, the practical changes might be more complicated than the change types discussed in the study. The author presented in-depth analysis of the proposed models referring to both limitations and implications as follows.

7.3.1. Limitations and Implications of the DSP-PROV Model

The proposed DSP-PROV model enables metadata maintainers to formally describe provenance of Description Set Profile and trace structural changes in metadata application profiles. From the perspective of standardization and interoperability of metadata application profiles, there may be limitations of the DSP-PROV model as follows.

The proposed model is applicable to those metadata schemas defined based on the Description Set Profile of the Singapore Framework of DCMI. There is still no world-wide standard of metadata application profiles. Not all metadata application profiles are defined following or can be easily converted into the DCAP.

Not all of those documents of metadata schemas are given in a formal description scheme like RDF. In the case of DPLA, those three versions of MAP are provided in pdf files and the change logs were in English. Therefore, the author had to manually process DPLA MAP in this study. In theory, it is possible to apply DSP-PROV model to any of MAPs defined based on the Singapore Framework of DCMI. However, in the reality, the application of the proposed model may be a case by case adaptation in the current environment.

DSP-PROV has functions to describe causal and derivation relationships. However, DSP-PROV does not have agent and temporal factors because they are not usually explicitly included in the schema documents. The author did not consider the temporal order of activities, i.e., temporal relationships between Activities. This is because this study aimed to develop a simple model with minimum set of Entities and Activities. The author considers that there is a need of further investigation on the extension of the proposed model which may handle these aspects.

In practice, on one hand, communities are developing their metadata application profiles to meet their specific needs. On the other hand, many communities have demands to mutually use their metadata across communities. Metadata mapping is a common solution to bridge the gap. Provenance description helps temporal mapping between metadata schemas. Mapping between metadata schemas in conventional environments are done manually, which fits to a conventional database-centric system environment of metadata. Metadata vocabulary registry services in the LOD environment such as Linked Open Vocabularies (LOV) can be used to support the mapping process by machines. The author discussed provenance of metadata application profiles, which are out of the scope of these registry services. These registry services can be extended to maintain metadata application profiles with the function of provenance tracking.

Consistent maintenance of metadata schemas for long-term is a natural demand for those metadata registries. Although metadata registries play crucial roles in the management and sharing of metadata terms, metadata vocabularies and metadata application profiles across communities and over time (Dunsire et al., 2012), metadata registries do not ensure metadata longevity. As those registries have to handle metadata schemas in various fields, they need a generalized model for metadata schema maintenance. Thus, the model proposed here is simplified for long-term maintenance of metadata in the LOD environment. This study is carried out in this stand-point. The method is still based on manual processing to convert the original documents into RDF. However,

the primary achievement of this study proved the applicability of the model to a metadata application profile used for a practical service. Proof of the concepts of DSP-PROV model in more automated process is left for future study.

Temporal interoperability is a vital facet for metadata interoperability (Li and Sugimoto, 2015). The use of provenance description has been proposed to obtain the temporal interoperability of metadata. Provenance of metadata application profiles enables to trace structural changes among different versions of metadata schemas. The achievements in this study would help metadata maintainers to maintain metadata application profiles, extend functions of metadata registries with provenance tracking, and audit errors in metadata mapping.

7.3.2. Limitations and Implications of the Vocab-PROV Model

The proposed Vocab-PROV model is defined based on a few primitive relationships (e.g., addition, deletion, and replacement) between pre-version and post-version of a metadata term. The model enables formal provenance description of metadata vocabularies in machine-processable form, which can improve maintainability of metadata vocabularies over time.

Compared to conventional maintenance of metadata terms that is the maintenance of documents of terms, the proposed Vocab-PROV model enables effective and automated tracking of change history of metadata vocabularies using simple and formal description scheme that is defined based on widely-used Web standards. Vocab-PROV provides a simple and formal scheme of provenance description of metadata vocabularies, which can work as the basis of automated maintenance of metadata terms and their vocabularies. Through keeping track of the changes in metadata vocabularies, it is helpful to audit inconsistencies in different versions of metadata vocabularies and facilitate the long-term maintenance of metadata.

On one hand, the Vocab-PROV model is simplified and applicable to metadata vocabularies in specific domains, such as medical subject headings, agricultural thesaurus. On the other hand, the practical changes of metadata terms can be more complicated than the primitive relationships discussed in the model. That is, Vocab-PROV is not a “one-fits-all” model that can be applied to track semantic changes in all vocabularies. For instance, the change history of the vocabularies represented in RDF is not easy to track, which raises up further issue of provenance description of RDF graphs.

7.4. Related Issues for Further Research

7.4.1. Context Construction with Provenance Information for Metadata Preservation

Provenance and context are crucial for understanding and justifying digital records. “Where records are concerned, documentation of provenance and context forms a basis for enhancing their transparency and thus for evaluating their trustworthiness” (Yeo, 2013). Modeling solutions for provision of contextual information for digital metadata objects need to be developed. Brocks et al. (2010) have proposed a generic context model aligned with the OAIS framework for digital preservation. They extended the OAIS information model by a specialized context information package. In their study, context of a digital object is defined as the representation of known properties associated with the digital object and the operations that have been implemented on the digital object. However, it is not easy to construct a complete, concise and unambiguous context for preserved metadata objects.

Provenance of the preserved metadata objects that describes the change history of the objects for the long term can be considered as part of context construction for metadata preservation. We consider that the context for preserved metadata objects may document the following information: the organizational and technical processes associated with their creation, ingestion, preservation, access and reuse; spatial and temporal information of the metadata objects; involved resources and agents; structural and semantic relationships with other related environments; etc. Presenting a timeline of events relevant to metadata objects and making connections between events and agents information (e.g., people, organization, software) may contribute to the context construction (Griner, 2008).

The context construction issue requires further research for the long-term usability of LOD on the Semantic Web, where a context can be represented as a Web resource and identified with an Internationalized Resource Identifier (IRI) that is a generalization form of URI and URL (Bao et al., 2010). With growth of metadata objects transferred and exchanged as LOD on the Web, formal descriptions of provenance and context can work well with Semantic Web technologies to assist reasoning with RDF and OWL.

7.4.2. Sharing Research Data with Provenance for Longevity of Research Data

Over the past few years, there have been national and international needs for archiving and preservation of research data. National and global agencies provide services for researchers to make research data accessible and reusable. The need for sharing research data without barriers and

across boundaries is increasing especially in the sciences. There are a lot of committees, services and projects to facilitate sharing of research data. For instance, National Science Foundation in the DataNet Program, Data Conservancy; the committee on Data for Science and Technology of the International Council for Science (CODATA); Australian National Data Service (ANDS); Research Data Alliance (RDA); Scalable Preservation Environments (SCAPE) project, etc. Moreover, research libraries are actively involved in the management of research data through provision of relevant services, e.g., services about data management consultation, data infrastructure development, data curation services. In practice, Collaborative Data Sharing Systems (CDSS) supporting to handle data provenance (e.g., provenance querying, data changes, schema mappings) have been developed so that scientists can specify whose data to trust (Karvounarakis, 2009; Green, 2009).

A model for formal representation of research and research data has been presented. The model was engineered as an ontology named Core Ontology for Scientific Investigations (COSI) to describe core entities of a research investigation focusing on provenance and contextual information (Brahaj, 2016). Provenance of research data explains where the data and research results came from, confirms data quality and permits replication experiments. Provenance is beneficial to scientific practice and reproducibility of research. Provenance for research data may include descriptions of employed equipment, mathematical and logical operations, oversight operations, and other process elements, which are necessary to make both the inquiry and its results clear and transparent to scientific colleagues and the interested public (Edward and Heather, 2014). High quality research data with trust needs long-term curation and preservation. Provenance of research data is essential to make research reproducible, enhance data reliability, and ensure reproducibility and attribution of research results.

Sharing data within and across disciplines needs research data management plan, license to acknowledge data rights, standardized citation mechanism to acknowledge data ownership, monitoring of the secondary usage of data, and provenance description over the creation and reuse of research data. Provenance over the lifecycle of research process can assist in the management of research data. The responsible agents, such as research project directors, research staffs that collect, process or analyze data, and their roles should be revealed as provenance to indicate attribution of research data. The research-related activities are also crucial components of provenance descriptions, for example, the generation, collection, process, analysis, storage, preservation, sharing, reuse, citation of research data, and so forth.

“A crucial part of making data user-friendly, sharable with long-lasting usability is to ensure they can be understood and interpreted by any user. This requires clear and detailed data description,

annotation and contextual information". Metadata and provenance over research lifecycle assist in the understanding and interpretability of research data (Van, 2011). From the research results of this study, the author has learned that keeping tracking of changes made to metadata over time is critical to metadata longevity. Provenance of research data over research process is beneficial to the long-term usability of research data. It will help researchers to identify trust and to participate in sharing research data without worries when provenance of research data is properly documented and provided. Meanwhile, high quality and trustable metadata are needed for sharing research data. Semantic Web technologies are also usable for sharing data, for instance, adding proper annotation of the research data.

Sharing research data with provenance can facilitate reuse of research data. Metadata plays a significant role in the sharing of research data for the long-term. Metadata is necessary for the management and long-term use of research data. The shared research data should be kept interpretable using their metadata and provenance. The creation of research data may also rely on metadata schemas and metadata vocabularies, and their longevity should be ensured for understanding research data.

8. Conclusion and Future Work

This chapter concludes the dissertation with brief summary of the research findings and suggestions on future research.

8.1. Summary of Research Findings

This dissertation addresses issues in longevity of metadata, especially long-term maintenance of metadata application profiles and metadata vocabularies. This study dealt with how to formally describe provenance description of metadata application profiles and metadata vocabularies, with emphasis on establishing models to trace their change history using Semantic Web technologies. The proposed models rely on several infrastructure issues in the practical Web environment.

The main contributions of this study are: (1) development of the DSP-PROV model for provenance description of metadata application profiles and application of the model to a case study, (2) development of the Vocab-PROV model for provenance description of metadata vocabularies, and (3) provision of examples of formal provenance description of metadata, especially considering both the structural changes in metadata application profiles and the semantic changes in metadata vocabularies. The provenance description of metadata application profiles and metadata vocabularies together can reveal the revision history of structural features and semantic features of metadata instances, which can help users to interpret metadata instances. These descriptions in machine-processable form on the Web can be traced using Semantic Web technologies.

The proposed DSP-PROV model is a generalized provenance model, which has only three primitive functions (i.e., deletion, addition and revision) to track structural changes in metadata application profiles. In general, formal provenance description in RDF has advantages in consistent management supported by machines and in exchanging provenance information in the Semantic Web environment. Formal provenance description based on the DSP-PROV model enables to explicitly describe changes of structural features in metadata schemas from different levels, such as Description Template level, Statement Template level, and Structural Constraint level defined in Description Set Profile of Dublin Core Application Profile.

The Vocab-PROV model is a general model for provenance description of metadata vocabularies to track the primitive changes of metadata terms between two consecutive versions of a metadata vocabulary, e.g., split and merge of metadata terms and revision of meaning of metadata terms. The model enables formal provenance description of metadata vocabularies, especially how metadata terms and term definitions change over time.

The main findings of this study are summarized into the following points: (1) provenance of metadata is crucial component for consistent maintenance of metadata, (2) formal provenance of metadata should be consistently recorded in machine-processable form on the Web, (3) the proposed DSP-PROV model and Vocab-PROV model enable metadata maintainers to keep track structural changes in metadata application profiles and semantic changes in metadata vocabularies, respectively, and (4) formal provenance description holds advantages over provenance description in natural languages. For instance, formal provenance description helps consistent maintenance of metadata over time; formal provenance description can be used to find errors in semi-provenance description that is recorded in a natural language.

Although there are some limitations of the models proposed in this study, the findings and results have contributed new sights to metadata longevity from the perspective of provenance. The author applied the W3C PROV for provenance description of metadata schemas to the metadata community. The proposed models can be used to assist metadata maintainers to manage metadata schemas and to extend functions of metadata registries with provenance tracking. Metadata registries should work with provenance for keeping metadata schemas as machine interpretable objects over time.

This study provides several fundamental issues for discussion for the future development of metadata longevity. Further research is needed to move the implementation and practice of metadata longevity forward, especially in the communities and institutions that create and manage digital collections for the long-term.

8.2. Suggestions for Future Work

There are many issues related to metadata longevity and provenance description of metadata in the digital environment and on Semantic Web, for instance, contextual information (Lee, 2011), provenance information, metadata interoperability, long-term maintenance of metadata vocabulary, persistence of URIs, and so forth. This section highlights a few of the future directions. The author does not provide in-depth discussion about these issues, which are left for further exploration in the research area of metadata longevity.

How to construct contexts is still not clear. Contextual information of metadata schemas is crucial for their longevity. Provenance is a kind of contextual information. In the curation of digital datasets, a particularly important set of contextual relationships are “provenance links”, which give answer to “where did a piece of data come from?” (Buneman et al., 2006). Every metadata schema designed for an application has its contexts that may or may not be described as a part of the schema. For instance, descriptions about selection process of metadata terms from standardized

vocabularies are helpful to know the context of the metadata schema and to correctly interpret metadata. It is important to include the context construction in the management process of metadata schema, which is a challenging issue to explicitly and consistently describe the contexts in the LOD environment. Management of contextual information for the longevity of metadata schemas is left for future study.

Activities to publishing provenance as LOD are useful to facilitate provenance interchange on the Web. The LOD activities and Semantic Web technologies raise up several new issues for long-term and consistent use of metadata, such as stability of identifiers. Persistence of namespaces and identifiers is a crucial issue for metadata that are shared and transferred in the Semantic Web environment. The URIs used to identify metadata schema instances, metadata terms, and metadata vocabularies should be kept persistent over time.

One of the remaining questions is how to make provenance descriptions beneficial to users. As discussed above, development of provenance-based Web services or applications or use cases for digital collections in memory institutions is a feasible way to reveal and provide provenance to users. For example, providing search service of metadata provenance, revealing metadata provenance in a visualization form together with related temporal information and responsible agents. Many visual analytics tools provide functions for visualization of provenance, such as Chimera, VisTrails, GraphTrail, the Graphical History Interface, TimeTravel interface, and so forth (Ragan et al., 2016). Building provenance repository in MLA community is another feasible option.

Another direction of future work is the use of provenance in metadata mapping. Metadata mapping tables are often created for merging two or more metadata datasets, metadata harvesting, federated search, and so forth. Metadata mapping table is also a crucial resource for long-time metadata maintenance. For instance, a vocabulary mapping table created for a metadata schema mapping is a metadata instance about the metadata schema mapping, e.g., conversion from an old schema to a new schema, and merger of two schemas. Provenance description for the mapping table should be provided to record the change history of metadata terms used in the schema(s). Metadata provenance is helpful to find errors and inconsistencies in the mapping among metadata schemas and the mapping among metadata vocabularies.

Provenance issues related to Web archiving are also future directions, such as provenance tracking of webpages, publishing provenance of web resources as LOD. Web archiving is important for keeping Web contents over time. The Web contents can be preserved through Web archives. Internet Archive project made efforts to record the history of the World Wide Web. For example, wayback machine service for saving webpages, Archive-It service for collecting and accessing cultural heritage on the Web. Provenance is necessary for the archived Web resources. Provenance

of the archived resources may include the content producers, URIs and timestamps of the archived Web resources, and how the Web contents evolved. Provenance information related to the different versions of a Web resource is useful to link the past with the present of the Web resource. Memento defines a framework to keep old URIs consistently usable using datetime negotiation (a variation on content negotiation) and TimeMaps. Memento framework enables to access the Web of the past. Coppens et al. (2011) have developed a digital long-term preservation archive that enables to produce and publish provenance as LOD.

Provenance is a multi-faceted research topic, which is not limited to identification of ownership or authorship but also refers to usage of Web technologies for description, management and preservation of provenance description in networked information environment. A suite of provenance tools has been developed by University of Southampton to deal with provenance that follows W3C PROV standard. The suite of software, libraries and services can be used to capture, visualize and store provenance in the Web environment (Moreau and Groth, 2013).

As a concluding remark of this dissertation, the author would like to look back at this study and address a few thoughts for the future. The author started this study from learning what is digital preservation and understanding problems in digital preservation from the viewpoint of metadata. She learned importance of provenance for longevity of digital objects and considered that metadata provenance is required for longevity of metadata since metadata are transferred as digital objects on the Web. Along with the study on metadata technologies and Web standards, the author deeply recognized that provenance of metadata schemas should be described in machine-understandable forms to fit to requirements of Web environment. By the implementation of the case study, the author learned that Semantic Web technologies such as OWL and SPARQL facilitate in the wide exchange of provenance among machines. She considers that provenance description should work with and take advantages of Semantic Web technologies for the description, capture, and reasoning of provenance among different systems.

The author conducted this study based on previous studies on digital preservation, metadata, and provenance. She noticed the potential usage of metadata provenance in digital libraries, digital archives and digital museums. She considered that provenance-based services in such memory institutions are upcoming. In the domain of Library and Information science, researchers are also exploring provenance of research data and provenance of Linked Data. These recent research activities bring more opportunities and issues for provenance research. The venue to move provenance research forward is still on-going. Overall, the insights presented in this dissertation would be useful for future research on provenance of metadata.

Acknowledgements

I would like to express my deepest thanks to my supervisor Professor Shigeo Sugimoto. He guided me patiently and gave me a lot of freedom to carry out my research project. He is an outstanding professor with rich experience and professional knowledge. He contributed a great deal to my Ph.D. study and carefully helped me check my dissertation. He shared his experiences about academic presentations and academic writing with me. He led me on the way to professional research. He gave me chances to go abroad for academic presentations and international communication. Thanks to him, I received many opportunities to be introduced to outstanding LIS research projects, researchers and communities by attending international conferences (DC, A-LIEP, ICADL, iConference, iPRES). He provided me a part-time research job to lighten my financial burden in the last year of my Ph.D. study. I have a social science education background and this Ph.D. project is not easy for me. So many thanks to my supervisor for his patience and kindness, and I could not have achieved the fruits of my research without his great guidance and kind help during the past few years. I learned rigorous academic attitude and a professional working style from him. I admire him and appreciated his guidance for my growth.

I would like to express my sincerest thanks to Professors Atsuyuki Morishima, Tetsuo Sakaguchi, and Mitsuharu Nagamori. They kindly guided me and shared their research experiences with me. They gave me a lot of good advice, constructive suggestions, and active feedback during graduate student seminar over the past few years. They are from different research areas, which helps to move this study forward after the exchange of ideas and thoughts. I appreciate their contributions to my Ph.D. study.

I also thank Prof. Marica Lei, Zeng working at School of Library and Information Science, Kent State University for her professional suggestions on my journal paper. I also want to express my thanks to Prof. Jian Qin at School of Information Studies, Syracuse University for her academic and career experiences that she shared with me. I also thank Dr. Thomas Baker a member of DCMI committee for giving me suggestions on metadata research. I also thank Prof. Jane Greenberg working at Metadata Research Center, Drexel University, who coauthored one of my conference papers. I also thank Dr. Bhuva Narayan, a senior lecturer working at University of Technology Sydney, who is warm hearted, helped me in English editing and friendly encouraged me a lot.

I also thank Dr. Tetsuya Mihara, Tsunagu Honma, Senan Kiryakos, Chiranthi Wijesundara, Winda Monika and other members in our metadata laboratory. We studied together and discussed about life and research during the graduate student seminar. Thanks go to their kind help. I cherish

the moments when we spent together during the seminar, conferences, lab parties during the study life in our big family. I wish all the best to their studies and career life in the future.

I also thank the financial support that I got from Education, Culture, Sports, Science and Technology (MEXT) of Japanese government and China Scholarship Council (CSC). Thanks to them, I was able to concentrate on my research without financial worry. Without the program, I would never have the chance to experience oversea study in my life.

I would like to thank the referees for their valuable comments on my publications accomplished during my Ph.D. research project. I would like to express thanks to the professors Tomoo Inoue, Shoichiro Hara, and Masao Takaku in the defense committee. I appreciate their coming to my defense and their reviews to my dissertation. I am deeply grateful to their questions, suggestions and comments, which helped me summarize the dissertation.

I also thank my master supervisor Prof. Chenying Li working at China Agricultural University Library. She patiently guided me during my master study and recommended Prof. Sugimoto to me for my Ph.D. study. And special thanks to my friends Hengjun Liu, Xi Han, HaiTao Yu, Bikun Chen …… who studied in Japan and China though I cannot list all of them here. We shared our Ph.D. life experience, exchanged views on study, and cheered each other up with hope.

At last, I want to express my thanks to my family. My parents, parents-in-law, husband, sister and brother give me a lot of moral encourage all the time. Thanks goes to my mother-in-law for taking care of my baby for almost one year. Thanks go to my parents for taking care of my baby for more than one and half years in Beijing though my mother has diabetes and my father has hypertension. Although they are not rich and do not have good educational backgrounds, they love me and help me when I need them. I thank my husband, who is a quite positive and responsible man. He took care of our daughter and my parents when I was studying in Japan. He helped me with family burden and stood behind me all the time. I appreciate what he did for me and for our family. I missed the growth of my daughter in her early age, and I hope I can provide her a good education environment in the future.

The years that I studied in Graduate School of Library, Information and Media Studies, University of Tsukuba are precious experience in my life. It is an exploration journey that helped me grow up in the academic community. Finally, I achieved to my goal of obtaining my Ph.D. degree. I appreciate what I learned and gained during my Ph.D. study. I am dreaming to be a professor in LIS in the future, which will be another long journey for me. My Ph.D. study comes to an end, but it is just a new transition point for me to work on my future career. I am looking forward to the new challenges in the upcoming career life. I am waiting to embrace the future.

References

Auer Sören, Theodore Dalamagas, Helen Parkinson, François Bancilhon, Giorgos Flouris, Dimitris Sacharidis, Buneman Peter, Kotzinos Dimitris, Stavrakas Yannis, Christophides Vassilis, Papastefanatos George, and Thiveos Kostas. (2012). Diachronic Linked Data: Towards Long-Term Preservation of Structured Interrelated Information. In WOD'12 Proceedings of the First International Workshop on Open Data, ACM, New York, NY, USA, pp. 31-39.

Anam Sarawat, Byeong Ho Kang, Yang Sok Kim, and Qing Liu. (2015). Linked Data Provenance: State of the Art and Challenges. In the 3rd Australasian Web Conference (AWC 2015), Vol. 166, pp. 19-28.

Brahaj Armand. (2016). Semantic Representation of Provenance and Contextual Information in Scientific Research. Ph.D. Dissertation. Humboldt-Universität zu Berlin.

Burgess Lucie C. (2016). Provenance in Digital Libraries: Source, Context, Value and Trust. Building Trust in Information. pp. 81-91. Springer International Publishing.

Baker Thomas. (2004). Maintaining a Vocabulary: Practices, Policies, and Models around Dublin Core. In 2004 Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2004), available at: http://dcpapers.dublincore.org/pubs/article/view/765

Baker Thomas. (2007). Dublin Core Metadata Element Set Detailed Change Description.

Baker Thomas, Pierre-Yves Vandenbussche, and Bernard Vatant. (2013). Requirements for Vocabulary Preservation and Governance. Library Hi Tech. Vol. 31, No. 4. pp. 657-668.

Baker Thomas, and Alistair Miles. Principles of Good Practice for Managing RDF Vocabularies and OWL Ontologies.

Buneman Peter, Adriane Chapman, and James Cheney. (2006). Provenance Management in Curated Databases. In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 539-550. ACM.

Batsakis Sotiris, Giaretta David, Gueret Christophe, van Horik Rene, Hoogerwerf Maarten, Isaac Antoine, Meghini Carlo, and Scharnhorst Andrea. (2014). PRELIDA D3.1 State of the Art Assessment on Linked Data and Digital Preservation. European Commission.

Brocks Holger, Alfred Kranstedt, Gerald Jäschke, and Matthias Hemmje. (2010). Modeling Context for Digital Preservation. Smart Information and Knowledge Management. pp. 197-226.

Baca Murtha. (2008). Introduction to Metadata. Getty Publications.

Bao Jie, Tao Jiao, McGuinness Deborah L., and Smart Paul. (2010). Context Representation for the Semantic Web. Web Science Conference, United States. 26-27 April 2010.

BBC. (2012). Provenance Ontology, available at: http://www.bbc.co.uk/ontologies/provenance (accessed 18 March 2014)

Ball Alexander. (2012). Review of Data Management Lifecycle Models.

BIBFRAME 2.0 Vocabulary List View. (2016). The Library of Congress website, available at: http://id.loc.gov/ontologies/bibframe.html (accessed 20 June 2016)

BIBFRAME 2.0 Specifications Notes. (2016). The Library of Congress website, available at: https://www.loc.gov/bibframe/docs/pdf/bf2-notes-june2016.pdf (accessed 20 June 2016)

Consultative Committee for Space Data System (CCSDS). (2012). Reference Model for an Open Archival Information System (OAIS). No. Recommended Practice, Issue 2, CCSDS 650.0-M-2, CCSDS Secretariat, Space Communications and Navigation Office, Space Operations Mission Directorate, NASA Headquarters, Washington, DC, available at: http://public.ccsds.org/publications/archive/650x0m2.pdf (accessed 29 March 2016).

Chao Tiffany C., Cragin Melissa H., and Palmer Carole L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An Empirically Derived Framework of Scientific Data Practices and Curatorial Processes. Journal of the Association for Information Science and Technology. Vol. 66, No. 3. pp. 616-633.

Chilvers Alison. (2002). The Super-Metadata Framework for Managing Long-Term Access to Digital Data Objects: A Possible Way forward with Specific Reference to the UK. Journal of Documentation. Vol. 58, Issue 2. pp. 146-174.

Ciccarese Paolo, Soiland-Reyes Stian, Belhajjame Khalid, Gray Alasdair JG, Goble Carole, and Clark Tim. (2013). PAV 2.0 - Provenance Authoring and Versioning Ontology. Journal of Biomedical Semantics. Vol. 4, No. 37, available at: http://www.jbiomedsem.com/content/4/1/37 (accessed 18 March 2014)

Chawuthai Rathachai, Takeda Hideaki, Wuwongse Vilas, and Jinbo Utsugi. (2016). Presenting and Preserving the Change in Taxonomic Knowledge for Linked Data. Semantic Web. Vol. 7, No. 6. pp. 589-616. http://doi.org/10.3233/SW-150192

Cuevas-Vicenttín Víctor, Ludäscher Bertram, Missier Paolo, Belhajjame Khalid, Chirigati Fernando, Wei Yaxing, Dey Saumen, Kianmajd Parisa, Koop David, Bowers Shawn, Altintas Ilkay, Jones Christopher, Jones Matthew B., Walker Lauren, Slaughter Peter, and Leinfelder Ben. (2015). ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance, available at: http://jenkins-1.dataone.org/jenkins/view/DocumentationProjects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html (accessed 30 March 2016).

Corrado Edward M., and Moulaison Heather Lea. (2014). Digital Preservation for Libraries, Archives, and Museums. Lanham: Rowman & Littlefield.

Dodds Leigh, and Ian Davis. (2012). Chapter 5. Data Management Patterns. Linked Data Patterns: A Pattern Catalogue for Modeling, Publishing, and Consuming Linked Data, available at: http://patterns.dataincubator.org/book/named-graphs.html (18 March 2014)

Dunsire Gordon, Corey Harper, Diane Hillmann, and Jon Phipps. (2012). Linked Data Vocabulary Management: Infrastructure Support, Data Integration, and Interoperability. Information Standards Quarterly. Vol. 24, No. 2/3. pp.4-13.

Digital Preservation Chapter. (2017). Digital Preservation: Putting it to Work. Vol. 700. Springer.

DPC Website. Risk and Change Management, available at: http://www.dpconline.org/handbook/institutional-strategies/risk-and-change-management.

Dappert Angela, Rebecca Squire Guenther, and Sébastien Peyrard. (Eds). (2016). Digital Preservation Metadata for Practitioners: Implementing PREMIS. Springer.

Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. (2013). PROV-O: The PROV Ontology, available at: http://www.w3.org/TR/prov-o/ (accessed 18 March 2014)

DCC. What is Digital Curation? http://www.dcc.ac.uk/digital-curation/what-digital-curation

DPLA MAP V3. (2013). Digital Public Library of America Metadata Application Profile, Version 3, available at: https://dp.la/info/developers/map/ (accessed 26 December 2016).

DPLA MAP V3.1. (2014). Digital Public Library of America Metadata Application Profile, Version 3.1, available at: https://dp.la/info/developers/map/ (accessed 26 December 2016).

DPLA MAP V4. (2015). Digital Public Library of America Metadata Application Profile, Version 4.0, available at: https://dp.la/info/developers/map/ (accessed 26 December 2016).

DCMI Metadata Terms. (2012). The Dublin Core Metadata Initiative website, available at: http://dublincore.org/documents/dcmi-terms/#terms-DCMIType (accessed 20 June 2016)

Dempsey Lorcan, and Heery Rachel. (1998). Metadata: A Current View of Practice and Issues. Journal of Documentation. Vol. 54, Issue 2. pp. 145-172.

Eckert Kai. (2012). Metadata Provenance in Europeana and the Semantic Web, available at: http://edoc.hu-berlin.de/series/berliner-handreichungen/2012-332/PDF/332.pdf (accessed 25 July 2014)

Eckert Kai. (2013). Provenance and Annotations for Linked Data. In 2013 Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2013), 2-6 September 2013, Lisbon, Portugal, pp. 9-18.

Flouris Giorgos, Roussakis Yannis, Poveda-Villalón María, Mendes Pablo N., and Fundulaki Irini. (2012). Using Provenance for Quality Assessment and Repair in Linked Open Data. In the 11th International Semantic Web Conference, 12 November 2012, Boston, Estados Unidos.

Factor Michael, Henis Ealan, Naor Dalit, Rabinovici-cohen Simona, Reshef Petra, and Ronen Shahar. (2009). Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. In Proceedings of TAPP'09 First Workshop on Theory and Practice of Provenance, San Francisco, CA, 23 February 2009, USENIX Association Berkeley, CA, USA.

Gil Yolanda, Miles Simon, Belhajjame Khalid, Deus Helena, Garijo Daniel, Klyne Graham, Misser Paolo, Soiland-Reyes Stian, and Zednik Stephan. (2013). PROV Model Primer, available at: https://www.w3.org/TR/prov-primer/ (accessed 29 March 2016).

Garijo Daniel, and Yolanda Gil. (2014). The OPMW-PROV Ontology, available at: http://www.opmw.org/model/OPMW/ (accessed 29 July 2014)

Griner Abigail R. (2008). Where's the Context? Enhancing Access to Digital Archives. Provenance, Journal of the Society of Georgia Archivists. Vol. 26, No. 1.

Greenberg Jane. (2005). Understanding Metadata and Metadata Schemes. Cataloging & Classification Quarterly. Vol. 40, No. 3-4. pp. 17-36.

Greenberg Jane. (2003). Metadata and the World Wide Web. Encyclopedia of Library and Information Science. pp. 1876-1888.

Green Todd J. (2009). Collaborative Data Sharing with Mappings and Provenance. Ph.D. Dissertation. University of Pennsylvania.

Grigoris Antoniou, Sotiris Batsakis, Antoine Isaac, Andrea Scharnhorst, José María García, René van Horik, David Giaretta, and Carlo Meghini. (2014). Analysis of the Limitations of Digital Preservation Solutions for Preserving Linked Data. PRELIDA EU Project, available at: http://prelida.eu/sites/default/files/PRELIDAD4.1.pdf

Grigoris Antoniou, Sotiris Batsakis, Antoine Isaac, Andrea Scharnhorst, David Giaretta, José María García, René van Horik, and Carlo Meghini. (2014). PRELIDA EU Project, Consolidated Roadmap.

Groth Paul, Gil Yolanda, Cheney James, and Miles Simon. (2012). Requirements for Provenance on the Web. The International Journal of Digital Curation. Vol. 7, No. 1. pp. 39-56. http://doi.org/10.2218/ijdc.v7i1.213

Hitzler Pascal, Krötzsch Markus, Parsia Bijan, Patel-Schneider Peter F., and Rudolph Sebastian. (2009). OWL 2 Web Ontology Language Primer. W3C Recommendation, available at: www.w3.org/TR/owl2-primer/

Harris Steve, Andy Seaborne, and Eric Prud'hommeaux. (2013). SPARQL 1.1 Query Language. W3C Recommendation, available at: http://www.w3.org/TR/sparql11-query/

Hein Stefan, and Schmitt Karlheinz. (2013). Risk Management for Digital Long-Term Preservation Services. In Proceedings of the 10th International Conference on Digital Preservation (iPRES 2013). 2-6 September 2013, Lisbon, Portugal, available at: http://phaidra.univie.ac.at/o:378059

Heery Rachel, and Patel Manjula. (2000). Application Profiles: Mixing and Matching Metadata Schemas. Ariadne. Issue 25, available at: http://www.ariadne.ac.uk/issue25/app-profiles (accessed May 27, 2016).

Halpin Harry, and Cheney James. (2014). Dynamic Provenance for SPARQL Updates using Named Graphs. In WWW'14 Companion Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7-11 April 2014, ACM, New York, NY, USA, pp. 287-288.

Hartig Olaf. (2009). Provenance Information in the Web of Data. In Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW 2009), Madrid, Spain.

Hartig Olaf, and Jun Zhao. (2010). Publishing and Consuming Provenance Metadata on the Web of Linked Data. International Provenance and Annotation Workshop (IPAW 2010). pp.78-90.

Hartig Olaf, and Jun Zhao. (2012, March 14). Provenance Vocabulary Core Ontology Specification, available at: http://trdf.sourceforge.net/provenance/ns.html (accessed 18 March 2014)

Hyland Bernadette, Atemezing Ghislain, Pendleton Michael, and Srivastava Biplav. (2013). Linked Data Glossary, available at: http://www.w3.org/TR/ld-glossary/ (accessed 20 June 2016)

Haynes David. (2018). Metadata for Information Management and Retrieval: Understanding Metadata and its Use, 2nd edition. Facet Publishing.

ISO 15489-1:2016. Information and Documentation – Records Management – Part 1: Concepts and Principles. (2016).

ISO 23081-1:2017. Information and Documentation – Managing Metadata for Records. (2017).

ISO 31000: 2009. Risk Management – Principles and Guidelines. (2009). International Organization for Standardization, Geneva, Switzerland.

Isaac Antoine, Waites William, Young Jeff, and Zeng Marcia. (2011). Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets, available at: https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/ (accessed 20 June 2016)

Isaac Antoine, and Summers Ed. (2009). SKOS Simple Knowledge Organization System Primer, World Wide Web Consortium (W3C), available at: http://www.w3.org/TR/skos-primer/

Jones Tanya Gray, Burgess Lucie, Jefferies Neil, Ranganathan Anusha, and Rumsey Sally. (2015). Contextual and Provenance Metadata in the Oxford University Research Archive (ORA). Metadata and Semantics Research. Vol. 544. pp. 274-285.

Javed Muhammad, Abgaz Yalemisew M., and Pahl Claus. (2014). Layered Change Log Model: Bridging between Ontology Change Representation and Pattern Mining. International Journal of Metadata Semantics and Ontologies, Vol. 9, No. 3, available at: http://doi.org/10.1504/IJMSO.2014.063137

Karvounarakis Grigorios. (2009). Provenance in Collaborative Data Sharing. Ph.D. Dissertation. University of Pennsylvania.

Kim Won. (2005). On Metadata Management Technology: Status and Issues. Journal of Object Technology. Vol. 4, No. 2. pp.41-48.

Kashyap Vipul, Christoph Bussler, and Matthew Moran. (2008). The Semantic Web: Semantics for Data and Services on the Web. Springer Science & Business Media.

Kunze John, Calvert Scout, DeBarry Jeremy D., Hanlon Matthew, Janée Greg, and Sweat Sandra. (2017). Persistence Statements: Describing Digital Stickiness. Data Science Journal. Vol. 16.

Kumar Sharma, Marjit Ujjal, and Biswas Utpal. (2013). Exposing MARC 21 Format for Bibliographic Data as Linked Data with Provenance. Journal of Library Metadata. Vol. 13, No. 2-3. pp. 212-229.

Kendall Elisa, Vit Novacek, Baker Thomas, and Miles Alistair. (2008). Principles of Good Practice for Managing RDF Vocabularies and OWL Ontologies. W3C editor's draft, available at: https://www.w3.org/2006/07/SWD/Vocab/principles (accessed 20 June 2016)

Knowlton Steven A. (2005). Three Decades since Prejudices and Antipathies: A Study of Changes in the Library of Congress Subject Headings. Cataloging & Classification Quarterly. Vol. 40, No. 2. pp.123-145. http://dx.doi.org/10.1300/J104v40n02_08

Lóscio Bernadette Farias, Burle Caroline, Calegari Newton, Greiner Annette, Isaac Antoine, Iglesias Carlos, Laufer Carlos, Guéret Christophe, Lee Deirdre, Schepers Doug, Stephan G. Eric, Kauz Eric, Atemezing A. Ghislain, Beeman Hadley, Bittencourt Ig Ibert, Almeida João Paulo, Dekkers Makx, Winstanley Peter, Archer Phil, Albertoni Riccardo, Purohit Sumit, and Córdova Yasodara. (2017). Data on the Web Best Practices, available at: http://www.w3.org/TR/dwbp/ (accessed 27 July 2016)

Lee Christopher A. (2011). A Framework for Contextual Information in Digital Collection. Journal of Documentation. Vol. 67, Issue 1. pp. 95-143.

Liu Jun. (2011). W7 Model of Provenance and Its Use in the Context of Wikipedia. Ph.D. Dissertation. Faculty of the Committee on Business Administration, University of Arizona.

Lebo Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, and Lemieux Victoria L. (2016). Building Trust in Information: Perspectives on the Frontiers of Provenance, Springer International Publishing, Switzerland.

Lebo Timothy, Sahoo Satya, McGuinness Deborah, Belhajjame Khalid, Cheney James, Corsar Daniel, Soiland-Reyes Stian, and Zhao Jun. (2013). PROV-O: The PROV Ontology, available at: http://www.w3.org/TR/prov-o/ (accessed 25 June 2016)

Leitch Matthew. (2010). ISO 31000: 2009 The New International Standard on Risk Management. Risk Analysis. Vol. 30, No. 6. pp. 887-892.

Lagoze Carl, Willliams Jeremy, and Vilhuber Lars. (2013). Encoding Provenance Metadata for Social Science Datasets. Communications in Computer and Information Science. Vol. 390, Springer International Publishing, pp.123-134. http://doi.org/10.1007/978-3-319-03437-9_13. Metadata and Semantics Research: 7th Metadata and Semantics Research Conference (MTSR 2013), Thessaloniki, Greece.

Li Chunqiu, and Sugimoto Shigeo. (2014). Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata. In 2014 Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2014), Austin, Texas, USA, 8-11 October 2014, Dublin Core Metadata Initiative, USA, pp. 147-156.

Li Chunqiu, Nagamori Mitsuharu, and Sugimoto Shigeo. (2015). Temporal Interoperability of Metadata: An Interoperability-Based View for Longevity of Metadata. In Proceedings of the 6th International Conference on Asia-Pacific Library and Information Education and Practice (A-LIEP 2015), Manila, Philippines, 28-30 October 2015, UP School of Library and Information Studies, Diliman, Quezon City, pp. 212-222.

Li Chunqiu, and Sugimoto Shigeo. (2016). Provenance Description of Metadata Vocabularies for the Long-Term Maintenance of Metadata. In Proceedings of the 7th International Conference on Asia-Pacific Library and Information Education and Practice (A-LIEP 2016), Nanjing, China, 3-4 November 2016, pp.150-164.

Li Chunqiu, and Sugimoto Shigeo. (2017). Provenance Description of Metadata Vocabularies for the Long-term Maintenance of Metadata. Journal of Data and Information Science. Vol. 2, No. 2. pp.41-55.

LCSH Introduction. (2016). The Library of Congress website, available at: https://www.loc.gov/aba/publications/FreeLCSH/lcshintro.pdf (accessed 20 June 2016)

Memento Protocol, available at: http://mementoweb.org/depot/native/dbpedia/

Mayernik Matthew S. (2016). Research Data and Metadata Curation as Institutional Issues. Journal of the Association for Information Science and Technology. Vol. 67, No. 4. pp. 973-993.

Mayernik Matthew S., Dilauro Tim, Duerr Ruth, Elliot Metsger, Thessen Anne E., and Choudhury G. Sayeed. (2013). Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation. Data Science Journal. Vol. 12. pp. 158-171.

Metadata Management White Paper. (2005). Metadata Management: An Essential Ingredient for Information Lifecycle Management.

Ma Xiaogang, Fox Peter, Tilmes Curt, Jacobs Katharine, and Waple Anne. (2014). Capturing Provenance of Global Change Information. Nature Climate Change, Nature Publishing Group. Vol. 4. pp. 409-413.

Missier Paolo, and Chen Ziyu. (2013). Extracting PROV Provenance Traces from Wikipedia History Pages. In EDBT'13 Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, Italy, ACM, New York, NY, USA, pp. 327-330. http://doi.org/10.1145/2457317.2457375

Missier Paolo, Dey Saumen, Belhajjame Khalid, Cuevas-Vicenttín Víctor, and Ludäscher Bertram. (2013). D-PROV: Extending the PROV Provenance Model with Workflow Structure. In TaPP'13 Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance, Lombard,

Illinois, 2-3 April 2013, USENIX Association Berkeley, CA, USA, available at: http://doi.org/10.1145/2457317.2457375

Meinhardt Paul. (2015). Versioning Linked Datasets: Towards Preserving History on the Semantic Web. Master Dissertation. University of Potsdam, available at: https://hpi.de/fileadmin/user_upload/fachgebiete/meinel/Semantic-Technologies/theses/Masterthesis-Meinhardt-2015.pdf (accessed 25 June 2016)

Moreau Luc. (2010). The Foundations for Provenance on the Web. Foundations and Trends in Web Science. Vol. 2, No. 2-3. pp. 99-241. http://doi.org/10.1561/1800000010

Moreau Luc, Paolo Missier, Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes. (2013). PROV-DM: The PROV Data Model, available at: http://www.w3.org/TR/prov-dm/ (accessed 18 March 2014)

Moreau Luc, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche. (2011). The Open Provenance Model Core Specification (v1.1). Future Generation Computer Systems. Elsevier B.V., Vol. 27, No. 6. pp. 743-756.

Moreau Luc, Li Ding, Joe Futrelle, Daniel Garijo Verdejo, Paul Groth, Mike Jewell, Simon Miles, Paolo Missier, Jeff Pan, and Jun Zhao. (2010). Open Provenance Model (OPM) OWL Specification, available at: http://openprovenance.org/model/opmo (accessed 18 March 2014)

Moreau Luc, Groth Paul, Cheney James, Lebo Timothy, and Miles Simon. (2015). The Rationale of PROV. Web Semantics: Science, Services and Agents on the World Wide Web. Elsevier B.V., Vol. 35. pp. 235-257. http://doi.org/10.1016/j.websem.2015.04.001

Moreau Luc, and Groth Paul. (2013). Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers.

Masó Joan, Closa Guillem, and Gil Yolanda. (2015). Applying W3C PROV to Express Geospatial Provenance at Feature and Attribute Level. Lecture Notes in Computer Science. Vol. 8628, Springer International Publishing, Switzerland, pp. 271-274. http://doi.org/10.1007/978-3-319-16462-5_31

Nagamori Mitsuharu, and Sugimoto Shigeo. (2004). A Metadata Schema Framework for Functional Extension of Metadata Schema Registry. In 2004 Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2004), available at: http://dcpapers.dublincore.org/pubs/article/viewFile/764/760/

Nilsson Mikael, Baker Thomas, and Johnston Pete. (2008). The Singapore Framework for Dublin Core Application Profiles, available at: http://dublincore.org/documents/singapore-framework/ (accessed 28 March 2016)

Nilsson Mikael. (2008). Description Set Profiles: A Constraint Language for Dublin Core Application Profiles, available at: http://dublincore.org/documents/dc-dsp/ (accessed 31 March 2016).

Nilsson Mikael, Miles Alistair J., Johnston Pete, and Enoksson Fredrik. (2009). Formalizing Dublin Core Application Profiles - Description Set Profiles and Graph Constraints. Metadata and Semantics. pp. 101-111.

Nilsson Mikael, Baker Thomas, and Johnston Pete. (2009). Interoperability Levels for Dublin Core Metadata, available at: http://dublincore.org/documents/interoperability-levels/

Omitola Tope, Christopher Gutteridge, and Nicholas Gibbins. (2011). Voidp: A Vocabulary for Data and Dataset Provenance, available at: http://www.enakting.org/provenance/voidp/ (accessed 18 March 2014)

Omitola Tope, Zuo Landong, Gutteridge Christopher, Millard Ian C., Glaser Hugh, Gibbins Nicholas, and Shadbolt Nigel. (2011). Tracing the Provenance of Linked Data using voiD. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS'11), 25-27 May 2011, Sogndal, Norway. ACM.

OCLC/RLG Working Group on Preservation Metadata. (2002). Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects, available at: www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf

OCLC/RLG Preservation Metadata Working Group. (2002). A Metadata Framework to Support the Preservation of Digital Objects. OCLC Online Computer Library Center, Dublin, USA.

OCLC/RLG Preservation Metadata Working Group. (2002). Preservation Metadata for Digital Objects: A Review of the State of the Art. OCLC Online Computer Library Center, Dublin, USA.

Pearce-Moses Richard. (2005). A Glossary of Archival and Records Terminology. pp. 317. Chicago: The Society of American Archivists, available at: http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf (accessed 18 March 2014)

Phipps Jon, Dunsire Gordon, and Hillmann Diane. (2015). Building a Platform to Manage RDA Vocabularies and Data for an International Linked Data World. Journal of Library Metadata. Vol. 15, No. 3-4. pp.252-264. http://doi.org/10.1080/19386389.2015.1099990

Poole Alex H. (2016). The Conceptual Landscape of Digital Curation. Journal of Documentation. Vol. 72, Issue 5. pp. 961-986.

Papastefanatos George. (2014). Challenges and Opportunities in the Evolving Data Web. In International Conference on Conceptual Modeling. In Advances in Conceptual Modeling. ER 2013. Vol. 8697. Springer International Publishing, Cham, pp. 23-28.

Patel Manjula. (2003). Ontology Servers and Metadata Vocabulary Repositories, available at: http://www.ukoln.ac.uk/metadata/agentcities/ (accessed 31 March 2016).

PREMIS Editorial Committee. (2015). PREMIS Data Dictionary for Preservation Metadata, version 3.0, available at: www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

PROV-FAQ. (2014). W3C Semantic Web, available at: https://www.w3.org/2001/sw/wiki/PROV-FAQ (accessed 8 April 2016).

Rothenberg Jeff. (1998). Ensuring the Longevity of Digital Information. Scientific American. Vol. 272, No. 1. pp. 42-7.

Ragan Eric D., Endert Alex, Sanyal Jibonananda, and Chen Jian. (2016). Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. IEEE Transactions on Visualization and Computer Graphics. Vol. 22, No. 1. pp.31-40.

RDA File Type. (2016). The Open Metadata Registry Website, available at: http://metadataregistry.org/vocabulary/show/id/98.html (accessed 30 June 2016)

RDA Sound Content. (2016). The Open Metadata Registry Website, available at: http://metadataregistry.org/history/list/vocabulary_id/93.html (accessed 30 June 2016)

Sahoo Satya S., and Amit Sheth P. (2009). Provenir Ontology: Towards a Framework for eScience Provenance Management, available at: http://corescholar.libraries.wright.edu/knoesis/80 (accessed 18 March 2014)

Schreiber Guus, and Raimond Yves. (2014). RDF 1.1 Primer. W3C Working Group Note, available at: http://www.w3.org/TR/rdf11-primer/

Sen Arun. (2004). Metadata Management: Past, Present and Future. Decision Support Systems. Vol. 37, No. 1. pp. 151-173.

Sharma Kumar, Ujjal Marjit, and Utpal Biswas. (2014). Generation with Provenance Tracking: A Review of the State-of-the-Art. International Journal of Scientific & Engineering Research. Vol. 5, No. 12. pp. 289-295.

Shaon Arif, and Andrew Woolf. (2008). An OAIS based Approach to Effective Long-Term Digital Metadata Curation. Computer and Information Science. Vol. 1, No. 2. pp. 2-16.

Shaon Arif. (2006). Long-term Digital Metadata Curation. In Proceedings of the UK E-Science All Hands Conference, Nottingham, United Kingdom, 18-21 September 2006, National e-Science Centre, UK, pp. 193-200.

Shaon Arif Bin Siraj. (2005). Long-Term Metadata Management & Quality Assurance in Digital Curation. Ph.D. Dissertation. University of Reading.

Sompel Herbert Van de, Sanderson Robert, Nelson Michael L., Balakireva Lyudmila, Shankar Harihar, and Ainsworth Scott. (2010). An HTTP-Based Versioning Mechanism for Linked Data. In Proceedings of Linked Data on the Web (LDOW 2010), Raleigh, North Carolina, USA, 27 April 2010, available at: http://arxiv.org/pdf/1003.3661v1.pdf

Sousa Renato Beserra, Cugler Daniel Cintra, Malaverri Joana Esther Gonzales, and Medeiros Claudia Bauzer. (2014). A Provenance-Based Approach to Manage Long-Term Preservation of Scientific Data. In 2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW 2014), 31 March 2014 - 4, April 2014, IEEE, Chicago, IL, pp. 126-133.

Theodoridou Maria, Yannis Tzitzikas, Martin Doerr, Yannis Marketakis, and Valantis Melessanakis. (2010). Modeling and Querying Provenance by Extending CIDOC CRM. Distributed and Parallel Databases. Vol. 27, No. 2. pp. 169-210.

Traczyk Tomasz, Włodzimierz Ogryczak, Piotr Pałka, and Tomasz Śliwiński. (Eds). (2017). Metadata in Long-Term Digital Preservation Chapter. Digital Preservation: Putting it to Work. Studies in Computational Intelligence. Vol. 700. Springer.

Tilmes Curt, Fox Peter, Ma Xiaogang, McGuinness Deborah, Privette Ana Privette, Smith Aaron, Waple Anne, Zednik Stephan, and Zheng Jin. (2013). Provenance Representation for the National Climate Assessment in the Global Change Information System. IEEE Transactions on Geoscience and Remote Sensing. Vol. 51, No. 11. pp. 5160-5168.

Tunnicliffe Sam, and Davis Ian. (2009). Changeset, available at: http://vocab.org/changeset/ (accessed 30 March 2016).

Völkel Max, and Groza Tudor. (2006). SemVersion: RDF-based Ontology Versioning System. In Proceedings of the IADIS International Conference on WWW/Internet 2006 (ICWI 2006), 5-8 October 2006, Murcia, Spain. pp.195-202.

Vermaaten Sally, Brian Lavoie, and Priscilla Caplan. (2012). Identifying Threats to Successful Digital Preservation: The SPOT Model for Risk Assessment. D-lib Magazine. Vol. 18, No. 9/10.

Van den Eynden Veerle, Louise Corti, Matthew Woollard, Libby Bishop, and Laurence Horton. (2011). Managing and Sharing Data: A Best Practice Guide for Researchers.

Vandenbussche Pierre-Yves, Atemezing Ghislain A., Poveda-Villalón María, and Vatant Bernard. (2017). Linked Open Vocabularies (LOV): A Gateway to Reusable Semantic Vocabularies on the Web. Semantic Web. Vol. 8, No. 3. pp. 437-452.

Wilson Thomas C. (2017). Rethinking Digital Preservation: Definitions, Models, and Requirements. Digital Library Perspectives. Vol. 33, No. 2. pp.128-136. https://doi.org/10.1108/DLP-08-2016-0029

Westbrooks Elaine L. (2005). Remarks on Metadata Management. OCLC Systems & Services: International Digital Library Perspectives. Vol. 21, No. 1. pp. 5-7.

White Paper Sun Microsystems. (2005). Metadata Management: An Essential Ingredient for Information Lifecycle Management.

W3C Provenance Incubator Group. (2010). Use Case Report, available at: http://www.w3.org/2005/Incubator/prov/wiki/Use_Case_Report (accessed 25 July 2014)

Yeo Geoffrey. (2013). Trust and Context in Cyberspace. Archives and Records. Vol. 34, No. 2. pp. 214-234.

Zhao Jun. (2010). Open Provenance Model Vocabulary Specification, available at: http://open-biomed.sourceforge.net/opmv/ns.html (accessed 18 March 2014)

List of Publications

Part 1: Peer-reviewed International Journal Papers

Chunqiu Li, Shigeo Sugimoto. (2017). Provenance Description of Metadata Vocabularies for the Long-term Maintenance of Metadata. Journal of Data and Information Science. Vol. 2, No. 2, pp. 41-55, https://doi.org/10.1515/jdis-2017-0007

Chunqiu Li, Shigeo Sugimoto. (2018). Provenance Description of Metadata Application Profiles for the Long-term Maintenance of Metadata Schemas. Journal of Documentation. Vol. 74, No. 1, pp. 36-61, https://doi.org/10.1108/JD-03-2017-0042

Part 2: Peer-reviewed International Conference Papers

Chunqiu Li, Shigeo Sugimoto. (2014). Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata. In *2014 Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2014)*, pp.147-156. Austin, Texas, USA, 8-11 October 2014, Dublin Core Metadata Initiative, USA. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3709/1932

Chunqiu Li, Mitsuharu Nagamori, Shigeo Sugimoto. (2015). Temporal Interoperability of Metadata: An Interoperability-based View for Longevity of Metadata. In *Proceedings of the 6th International Conference on Asia-Pacific Library and Information Education and Practice (A-LIEP 2015)*, pp.212-222. Manila, Philippines, 28-30 October 2015.

Chunqiu Li, Shigeo Sugimoto. (2016). Provenance Description of Metadata Vocabularies for the long-term maintenance of Metadata. In *Proceedings of the 7th International Conference on Asia-Pacific Library and Information Education and Practice (A-LIEP 2016)*, pp.150-164. Nanjing, China, 3-4 November 2016.

Shigeo Sugimoto, Chunqiu Li, Mitsuharu Nagamori, Jane Greenberg. (2016). Permanence and Temporal Interoperability of Metadata in the Linked Open Data Environment. In *2016 Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2016)*, pp.45-54. Copenhagen, Denmark, 13-16 October 2016. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2016/paper/view/430/473

Chunqiu Li, Shigeo Sugimoto. (2017). Metadata-Driven Approach for Keeping Interpretability of Digital Objects through Formal Provenance Description. In *Proceedings of the 14th International Conference on Digital Preservation (iPRES 2017)*, Kyoto, 25-29 September 2017, 10 pages.

Part 3: Workshop Position Paper without Peer-review

Shigeo Sugimoto, Chunqiu Li. (2017). Metadata Issues in Long-term Management of Data and Metadata. Information Science to Data Science: New Directions for iSchools Workshops at iConference 2017, Wuhan, Hubei, China, 22-25 March 2017.